# CONSTRUCTION OF A RATING SCALE FOR WRITING ASSESSMENT IN AN EFL CONTEXT

JYI-YEON YI

PH.D.

THE UNIVERSITY OF EDINBURGH

2006

*To my parents-in-law and parents*
*whose deep love for me*
*can never be measured by any rating scale*

# TABLE OF CONTENTS

## ABSTRACT

As the need for writing skills came to be seen as more and more important, the English Writing course for foreign language high schools in Korea was established and the first textbooks were published in 1997, within the 7th national curriculum. However, there are no suggested rating scales that can be used to assess this course, nor are published rating scales appropriate. Given this context, the purpose of this thesis is to construct a rating scale to assess this course using a data-based approach, and to validate the scale in terms of practicality, reliability and validity.

After obtaining three hundred and ninety scripts for two kinds of writing tasks carried out by Kwacheon foreign language high school students, I asked three English teachers at different high schools to assess them, to keep diaries for the investigation of their rating process and to do a think-aloud using both the subjective holistic scoring and the FCE rating scale. From these procedures, it was found that both rating schemes had problems in terms of validity, and that the FCE rating scale in particular was not appropriate for the Korean situation. Furthermore, a questionnaire study on one hundred and four English teachers at foreign language high schools across Korea revealed that they found it desirable to develop a rating scale for these students. I therefore attempted to develop a rating scale for this context. I coded all the scripts according to a coding scheme consisting of eighty-two coding categories that I had developed on the basis of the scripts themselves, the definition of writing ability implicit in the course, and my literature review, and statistically analysed the data to find the features discriminating between neighbouring bands. Finally, I constructed the first version of the rating scale, which was an analytic scale of six bands including three main assessment categories (Accuracy, Fluency and Organisation). The scale was then revised according to the feedback from the three raters through diary studies and questionnaire studies. The revised rating scale was investigated in terms of practicality, reliability and validity using eight quantitative or qualitative research methods, and found to be generally satisfactory except in certain fairly minor respects. It was also found to be different from the FCE rating scale. I will argue that this revised scale is valid for the given context, that it reflects the characteristics of Korean students' English writing, rather than making inappropriate assumptions based on theories of writing development,

that it has removed most of the problems found with the existing scales, and that it differentiates between bands in terms of qualitative as well as quantitative aspects.

## ACKNOWLEDGEMENTS

I would like to begin by thanking God first, for having always been with me, allowing and guiding me to do this work through His grace, and enabling me to realise how much I am loved by Him.

I am also deeply indebted to my supervisors, Mr. Brian Parkinson, Professor Alan Davies and Dr. Cathy Benson. Over the years they have been everything that supervisors should be: unsparing in both their encouragement and criticism; demonstrating a researcher's unceasing passion for research; and showing me how genuinely supportive teachers/supervisors can be of their students.

My grateful thanks also go to Dr. Glenn Fulcher and Mr. Eric Glendinning, for their careful reading of my thesis and detailed, integrative and insightful feedback on it as examiners for my viva. Their comments widened my theoretical and intellectual understanding and knowledge of both language testing and writing, in ways that will be extremely valuable to my future work. In addition, they taught me a lot about how an assessor's sympathetic consideration of the test-takers' position can help them demonstrate their real abilities in test situations.

My heartfelt thanks go to my supervisor in Korea, Professor Kyung-ja Park. She has been a constant source of advice and encouragement throughout my work on this thesis.

I would like to offer a word of special thanks to the participants in this research too: the many unknown Korean high school students who completed the writing tasks and questionnaires; the forty-two Korean teachers of English at foreign language high schools across Korea who answered a questionnaire study; Yong-kyun Choo and Seok-hwan Cheong who helped me obtain writing samples for this research; and last, but by no means least, my patient raters, Moon-cheol Kim, Won-kyung Choi and Sun-yang Ryu. This research would not have been possible without their help.

I also received a tremendous amount of support from my other colleagues, who assisted in various ways: Sae-bom Cheon with statistical analysis; Yu-kyung Choo with quality control in coding; Lou Leask with editing this thesis for me, a non-native speaker of English; and many of my colleagues in Korea who followed my progress throughout this research.

I particularly wish to record my appreciation for my parents-in-law, parents, sisters-in-law and younger brother, as well as my Korean friends in Edinburgh, for their prayers, concern and substantial support to me and my family during our stay in Edinburgh.

And finally, I must thank my immediate family: my two beloved children, Si-hyun and Seong-bhin, whose healthy and cheerful progress over the last four years has encouraged me in my work; and my husband, for his unfailing pleasantness and faith in me, and constant willingness to offer help and advice despite working hard on his own Ph. D.

Whilst I was able to complete this thesis thanks to the help of everyone mentioned above, any deficiencies it may contain are entirely my responsibility.

## AUTHOR'S DECLARATION

I declare that this thesis is my own work. The thesis has not been submitted for any other degree or professional qualification.

**Jyi-yeon Yi**

## LIST OF TABLES AND FIGURES

## ABBREVIATIONS

| | |
|---|---|
| ACTFL | American Council on the Teaching of Foreign Languages |
| ALTE | Association of Language Testers in Europe |
| ANOVA | Analysis of Variance |
| ASLPR | Australian Second Language Proficiency Ratings |
| CAE | Certificate in Advanced English |
| CPE | Certificate of Proficiency in English |
| EAP | English for Academic Purposes |
| ESP | English for Specific Purposes |
| FCE | First Certificate in English |
| FLHSK | Foreign Language High Schools in Korea |
| FSI | Foreign Service Institute |
| G-study | Generalisability study |
| IELTS | International English Language Testing System |
| IRT | Item Response Theory |
| ISLPR | International Second Language Proficiency Ratings |
| MBTI | Myers-Briggs Type Indicator |
| MTMM | Multitrait-multimethod |
| OPI | Oral Proficiency Interview |
| RS1 | First version of the new Rating Scale |
| RS2 | Revised version of the new Rating Scale |
| TEEP | Test in English for Educational Purposes |
| TWE | Test of Written Examination |
| TWP | Test of Writing Proficiency |
| UCLES | University of Cambridge Local Examination Syndicate |

# PART A.

# BACKGROUND

# CHAPTER ONE. INTRODUCTION

## 1.1 Introduction

This chapter will introduce the background, purpose, significance and research questions of this study. These will be followed by an introduction to the English educational situation in Korea and the English Writing course at high schools, and a discussion of why existing rating schemes are not satisfactory and a new rating scale needs to be developed for this course. Then, I will overview the chapters in this thesis and define some of the terms used in it.

## 1.2 Background to the study

Since working as an English teacher at Chung-shin girls' middle school in Seoul, Korea in 1997 and 1998, I have had an interest in language testing, in particular rating scales for both writing and speaking skills.

Afterwards, I had an opportunity to teach a basic course of English for freshmen at a college in Korea in 1999 and 2001. I had the students do continuous writing of one page about once a week. Every time that I tried to assess them, however, I found it very difficult to rate them. I believe this was partly because I was a non-native speaker of English and partly because I had no standards to use for assessment.

It was not until 2001, when I met and talked with Dr. Choi, Kyunghee,[1] then a lecturer in in-service training for English teachers at high schools in Korea, that I determined to develop a rating scale for writing assessment. I found out from talks with her that teachers at high schools were at a loss as to how to assess students' writing, and this gave me the incentive to undertake this research.

## 1.3 Purpose of the study

Given the increasing interest and need in writing, a new English Writing course for Korean high school students has been taught since it was introduced as part of the 7th national curriculum in Korea in 1997 (see section 1.6 for detailed discussion). Nonetheless, there is no rating scale suggested in the curriculum for teachers to use to assess the course. They have three choices available for the assessment in this

[1] She is a professor at Hangyang women's college in Seoul, Korea now.

situation: subjective holistic scoring (see section 1.8 for the definition of this term), using teacher's own scales and using published rating scales. None of these are, however, deemed to be satisfactory (I will discuss this point in section 1.6). Therefore, this study aims to construct a rating scale for assessment of the course through a data-based approach, and to examine its practicality, reliability and validity.

## 1.4 Significance of the study

The rating scale to be developed for this study is intended for the English Writing course, which is only taught at one of the various kinds of high school in Korea, Foreign Language High Schools in Korea (FLHSK).[2] There are over twenty FLHSK. Therefore, it might be considered that the new rating scale would be for a very small section of all high schools in the country. I expect, however, that the development of the rating scale will be significant in its own right.

First of all, the rating scale and the process of developing the scale could be used as a reference point when rating scales for the other schools are to be developed in the future. Since English writing skills are now so important for communication in international societies that they are being assessed in standardised tests such as TOEFL, and since performance assessment plays an ever greater role in classroom testing situations in Korea, the course could be extensively taught to students in other types of high school in Korea in the future, and rating scales appropriate for this situation will be required. In this case, I hope that this study will be regarded as valuable data.

Second, I hope that the rating scale will have a positive backwash on the teaching and learning of English writing in Korea. According to a private interview with an English teacher in the country, it seems that even though the English Writing course has been established, some teachers who are in charge of the course do not run it as required. I suppose that the reasons for this may be twofold: (1) since writing skills are not yet included in any university entrance examinations in Korea, these skills are not regarded as important; (2) since teachers are not given any useful rating schemes for the course, they are given little guidance on how to assess students' writing. Given that students and their parents are very sensitive to high

[2] See section 1.6 for more details.

school grades, which are important in securing a place at university/college (see section 1.6.1 for detailed discussion), the lack of a recognised rating scheme inhibits both fair assessment and proper delivery of the course. As university entrance examinations cannot be changed within a short period to include writing skills, this study is unlikely to have an immediate effect on them. However, the scale could address the second issue. If teachers have a helpful rating scheme, they would be more likely to run the course and assess students as required, which would play a significant role in leading students towards the intended goals of the course and helping them become aware of the importance of writing.

I hope that this study will be the first step towards these objectives, and that it will be meaningful.

## 1.5 Research questions

The research questions in this study are as follows:

(1) Is there a need to develop a rating scale for English Writing course in FLHSK?

(2) How can a rating scale with reasonable face validity be constructed?

(3) Is the rating scale thus proclaimed practical, reliable and valid?

The first research question covers the issue of why it is desirable to develop a rating scale for FLHSK, even though there are other rating schemes available. The second question deals with the methodology used to develop the rating scale, and the third question looks at the characteristics of the rating scale. These three questions cover the need to develop a rating scale for Korean students, the procedure for developing the scale, and its validation.

## 1.6 The problem

### 1.6.1 The education system in Korea

Since 1954, the education system in Korea has consisted of six years of primary school, three years of middle school, three years of high school and either four years of university or two years of college. Primary and middle school education are compulsory, but high school and university (or college) education are not, so students need to take an entrance examination to get into higher education.

There are various kinds of high school in Korea, such as general academic high schools, high schools for science education, FLHSK, high schools for commercial education, high schools for industrial education and art high schools. Before students choose which kind of school they wish to attend, their achievement during middle schools is considered. Those who achieve good marks can go to schools such as general academic high schools, FLHSK or high schools for science education, whilst those with lower marks go to schools such as the high schools for commercial education or the high schools for industrial education. (Entry into art high schools requires artistic talent as well as achievements during middle schools). In addition, those who wish to enter FLHSK and science education high schools have to take entrance examinations, which are devised by type of school concerned. Since the admission standards for examination by these two kinds of schools are very demanding, the academic level of those who gain entry through these examinations is high, although there is some variation depending on the school.

One general way for high school students to advance to university or college after graduating from high school is to take a university entrance examination. This is a standardised test carried out at national level. Since most high school students in Korea wish to enter university/college, the competition is extremely fierce. Therefore, all the education in high schools is entrance examination-oriented, and all teachers, students and their parents are concerned with and sensitive to the test. In addition, since not only the score on the standardised test but also general academic achievement during the three years at high school are considered for admission to universities/colleges, students and their parents are sensitive to the grades achieved during high school.

In this four-step-educational system, all schools except for universities/colleges follow the Korean national curriculum. This national curriculum specifies the aim, content, administration and testing of school education. In addition, textbooks produced by various publishers for this curriculum need to be authorised by the Ministry of Education. This curriculum was first established in 1954 and was revised six times between then and 1997, when the 7th national curriculum was introduced (Lee *et al.*, 2001). This 7$^{th}$ curriculum is still being implemented, and will be discussed in detail in the next section, with a particular focus on the English courses

in high schools.

### 1.6.2 The 7th national curriculum for English in high schools

According to the 7th national curriculum for English in high schools, the focus of English education is as follows: learner-centred education, considering individual differences; communicative ability-oriented education; activity/task-based education; logic and creativity-centred education; education contributing to national development and globalisation; living English and business English-focused education; specification of clear criteria for achievement; establishment of content through a systematic approach; settlement and activation of open education.

This national curriculum specifies several courses for high schools: General English I, General English II, English Reading Comprehension, English Conversation and English Writing. The first two are general courses that cover all four English skills, while the others are more specialised courses as their names indicate. Therefore, only the first two are covered at general academic high schools, high schools for commercial education, art high schools, high schools for industrial education and high schools for science education, but all the courses are taught at FLHSK. English Grammar and British/American Culture courses are also taught at FLHSK, even though there are no authorised textbooks for them.

The courses for the 7th national curriculum are different from those in the 6th national curriculum. Under the 6th national curriculum the English courses for high schools were General English I, General English II, English Reading Comprehension, English Conversation and Business English. The first three courses were offered in all kinds of high school, while the last two were for FLHSK only. It is worth noting that the Business English course in the 6th national curriculum was replaced by the English Writing course in the 7th one. In fact, English writing skills are also tackled in the General English I and General English II courses, although these skills are taught only to a limited extent. In addition to this, English Writing has been introduced as a separate course.

By way of introduction to the curriculum, the English Writing course is limited to FLHSK at the current stage (Lee *et al.*, 2001), though it may possibly and hopefully be extended to other high schools in the future, including general academic

high schools. In the next section, I will give a detailed introduction to the English Writing course, which is associated with this study.

### 1.6.3 English Writing course

The goal of this course is outlined in the course guidelines as follows:

> The English Writing course is the course whose focus is on developing the learners' ability to express their own thoughts and feelings in written English using the topics and language that have been taught in basic English courses. Whilst the goal of basic English courses is to help students express their thoughts in written language and communicate with target language users, this English Writing course is to help students not only express their thoughts in written language coherently and in an organised manner but also develop the ability to write accurately and fluently.
>
> The syllabus of this course should be planned to help students develop the ability to do the following things: to take notes regarding the theme of texts which they read or listen to; to write down the information, which they obtain during conversation, in an organised manner; to keep a diary and write informal or formal letters or emails; to write a description of their impression after reading literary works; to write reports on topics such as geography, history, arts, and education. Through these, this course aims at helping students to have confidence that they can express as well in written language as they can express in spoken language. To this end, it is desirable to help them do free writing rather than guided writing.
>
> (Lee *et al.*, 2001: 195)[translated into English from Korean]

In summary, the aim of the English Writing course is to encourage students to express their thoughts and feelings in an organised, accurate and fluent manner for effective communication across various genres (Lee *et al.*, 2001). In order to do this, this course aims to provide students with opportunities to produce various genres of writing, from private writing such as diaries, notes and private letters to formal, academic and business-related writing such as formal essays, agenda, reports, business letters, instructions and resumes.

The general guidelines for this course include the following directions:

- To make the course related to general academic English courses;
- To make a learner-centred syllabus which helps them be motivated;
- To use various audio-visual materials to help their learning;
- To teach plain and comprehensible content, taking into consideration the level of students' knowledge;
- To apply learner-centred teaching methods appropriate for teaching;

- To help students develop the ability to express themselves on practical topics;
- To help get students interested in writing through teaching how to write rather than through grammar-oriented teaching;
- To encourage students to apply logical ways of thinking that reflect the way native English is spoken;
- To encourage students to self-initiate learning, establishing the goals of learning and written topics by themselves.

(Lee *et al.*, 2001) [translated into English from Korean]

Several kinds of authorised textbooks for this course have been published according to these guidelines, and each school is allowed to choose any of them. Students at FLHSK are taught the course for two sessions per week. The point at which students will be offered the course during the three-year period at high school is at each school's discretion.

### 1.6.4 Assessment for the English Writing course

#### 1.6.4.1 Suggestions for the assessment of the course in the guidelines

When students are taught writing in the course according to the curriculum, they need to be assessed once or twice in a semester and ratings from the assessment constitute the grade that each student is given for the course and the grade is reflected in his/her achievement during high schools which is considered for admission to universities/colleges. Since Korean students, their parents and teachers at schools are university entrance exam-oriented as discussed in section 1.6.1, they are concerned about and sensitive to achievement/grade for each course during the schools. It is evidently the case with the performance-related course such as this writing course. In this context, with regard to assessment for the course, the guidelines set out the overall direction as follows:

- To assess through incorporated testing (where required, sometimes through discrete testing);
- To assess writing ability in various aspects, such as organisation, vocabulary, cohesion, spelling and expression;
- To assess free writing (and sometimes guided writing, when required);

- To assess the ability to write complete paragraphs and texts, rather than the ability to write accurately at sentence level;
- To give weight to fluency rather than accuracy at the onset of the course, and subsequently strive for a balance between fluency and accuracy.

(Lee *et al.*, 2001)[translated into English from Korean]

These suggestions are not very concrete, only giving general principles. In addition, the guidelines suggest subjective holistic scoring as a means of assessment in the course.

#### 1.6.4.2 Lack of suggested rating scales

As mentioned in section 1.6.4.1, the guidelines suggest the general direction of the assessment of this course, which is to assess by subjective holistic scoring, but they do not offer practical yardsticks that teachers teaching the course are recommended to use. They simply take one holistic scale and one analytic scale as examples for the explanation of rating scales. This leaves teachers with three options for assessing the course: (1) subjective holistic scoring; (2) using their own rating scales; or (3) using published rating scales.

#### 1.6.4.3 Inappropriateness of the three choices: subjective holistic scoring, using one's own rating scales and using published rating scales

The first choice, subjective holistic scoring, means that raters assess test-takers' work according to their own judgement based on an overall impression of performance, without using any rating scales (see section 1.8 for the discussion of this term). This kind of scoring has been criticised in terms of both the reliability and validity of the scoring. It generally does not ensure consistency, even intra-rater reliability, let alone inter-rater reliability. This is because rating is susceptible to influence from the following variables: the raters' ability to judge all the work they assess consistently; the extent to which raters compare the quality of the work being assessed with other pieces; the effects of tiredness on raters over the course of the rating session, and so on.

As for validity, it is likely that features such as content, organisation, vocabulary,

grammar and mechanics,[3] which are usually assessed in writing assessment, are not adequately measured, and that other qualities, such as handwriting and effort will be focused upon instead. In brief, where writing is assessed by subjective holistic scoring, scoring is likely to be less reliable or valid. This view is supported by both my questionnaire survey of English teachers and an empirical study, whose results are presented in section 5.2.5 and section 6.2.3 respectively. The fact that subjective holistic scoring is problematic in terms of validity should not be overlooked when considering the accountability measure of the guidelines, in that invalid assessment through subjective holistic scoring does not contribute to achieving the aim of the curriculum.

As a second choice, teachers may use their own rating schemes. If the schemes are well designed, they can work well. However, a questionnaire survey showed that Korean teachers of English were concerned that they may also result in the same problems as subjective holistic scoring (see section 5.2.5 and Appendix 3).

English teachers have a third choice: using one of many published rating scales. Published rating scales for continuous writing assessment include: the International English Language Testing System (IELTS) scales (www.ielts.org), scales for the Test of Written Examination (TWE) of the ETS (www.toefl.org), the American Council on the Teaching of Foreign Languages (ACTFL) guidelines (ACTFL, 1987), International Second Language Proficiency Ratings (ISLPR) (http://www.gu.edu.au/centre/call), scales for the Certificate in Advanced English (CAE) (CAE Handbook, 2001), Certificate of Proficiency in English (CPE) scales (CPE Handbook, 2002), scales for the Test in English for Educational Purposes (TEEP) (Weir, 1990), the ESL Composition Profile for the Michigan test battery (Jacobs *et al.*, 1981), the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1990, 1991b), and the FCE scales (FCE Handbook, 2001). There is also a scale published in Korea for the Test of Writing Proficiency (TWP, www.teps.or.kr).

None of these rating scales appear to be satisfactory for assessing the English Writing course, for the reasons outlined below.

---

[3] These categories are taken here as examples of qualities for which writing ability may be assessed, as they are common categories for writing assessment. However, this does not mean that they are standards which must be included in every rating scheme.

#### 1.6.4.3.1 Each rating scale has a specific context for which it is valid

In the past, the chief concern in performance tests was reliability, but this has now been replaced by validity (see sections 3.2 and 3.3.1 for further discussion), that is, whether a test or rating scale assesses what it intends to assess (see section 3.3.1.2 for a discussion of validity). What to assess in turn depends on the purpose of testing in a given situation, and other factors such as the level and need of the test-takers. Some rating scales aim to measure general proficiency in writing, others aim to measure what test-takers have been taught in a specific course. Some are concerned with grammatical accuracy at sentence level and content, others intend to assess overall organisation and ability to produce a coherent discourse. Some measure only linguistic ability, others also measure content-related knowledge. In other words, the construct of writing ability is different in each case. In addition, some are for nationwide or worldwide-standardised high stakes testing, others are for small-scale low stakes testing. Many rating scales are devised for specific tests, and all are developed to be valid for their own specific situation. Therefore, if they were used for different purposes and different situations from those for which they were developed, they would become invalid. This does not mean that they are invalid in their own right, but that their validity is situationally limited.

Unfortunately, however, none of the rating scales listed above was developed for the situation in Korea. There are none which can assess the very construct of writing ability implicit in the course. If the aforementioned scales were used for this specific situation, the ratings derived from them would not be valid in light of the construct of writing ability and purpose in this context. Therefore, it is not appropriate to use them in this context. Just as each of those scales has its own context, for which it was specifically developed despite the existence of many other rating scales, the context in Korea is different from the contexts to which these scales apply.

#### 1.6.4.3.2 Published rating scales fail to capture Korean students' unique features in English writing

Due to L1 transfer or cross-linguistic influence (Benson, 2002; Odlin, 1989) and other reasons such as input from teachers and textbooks, Korean learners of English

may show unique features in their English writing. For example, they may tend to use specific vocabulary and constructions, and their rhetoric (Choi, 1988; Eggington, 1987) may also be idiosyncratic. In addition, they may translate Korean literally into English and produce awkward expressions. These features should be reflected in a rating scale so that it can help raters assess students' writing and give the students more concrete information about the weaknesses/strengths of their English writing. At present, as the published rating scales aim to assess the entire pool of L2 learners, these specific characteristics in Korean students' English writing are not likely to be saliently incorporated into them.

#### 1.6.4.3.3 Published rating scales have problems because they have been developed *a priori*

Many published rating scales have been constructed *a priori*. Accordingly, they have problems due to the process of development. As Clark (1985) points out in his definition of rating scales, they are "descriptions of expected outcomes or impressionistic etchings of what proficiency *might* look like as one moves through hypothetical points or levels on a developmental continuum" (p.348). That is, "they represent teachers' conceptions of what students are expected to be able to do as opposed to what they actually do" (Scarino, 1995: 35). These scales might look logical, but as they have been developed *a priori* and are not based on learners' empirical data, "with regard to the basis for ordering descriptors as levels, a major difficulty arises from their linear nature" (Scarino, 1995: 36). Other problems resulting from *a priori* development are described in section 4.5.2.

To summarise, assessment of the course that appears to be both fair to students and their parents and valid for the purpose of the aim of the national curriculum needs to be made. However, aforementioned three scoring schemes (subjective holistic scoring, using teachers' own scales and using published rating scales) that are available are not satisfactory in assessing the course and test-takers in question. The solution proposed is to develop a new rating scale specifically for this course and these students, using a data-based approach.

## 1.7 Overview of the chapters

This study consists of two parts: Part A, Background, and Part B, The study. Part A consists of four chapters. Chapter One deals with the need for this study, its goals, and other introductory matters.

In Chapters Two to Four, I review the literature on writing ability, writing assessment and scoring criteria respectively. These three topics relate to three issues which must be considered before trying to develop a rating scale: what the scale is intended to measure, the assessment context in which it is to be used, and what kind of scale it needs to be. The first issue, "what is to be measured" addresses the construct: is it the ability to write correctly and fluently that is to be measured, or the ability to use the sources to support one's opinion, or something else? The second issue includes the kind of writing assessment to which the scale is going to be applied: impromptu assessments, portfolio assessments or another type of writing. Finally, "what kind of scale" refers to the method employed for the measurement: is it a holistic scale, an analytic scale, a primary trait scale or a multiple-trait scale, and is it a data-based scale or an *a priori*-developed scale? Since these three issues need to be decided in advance,[4] I review previous studies relating to these issues over three chapters.

In Part B I discuss how I developed the rating scale. Part B consists of five chapters, from Chapter Five to Chapter Nine. Chapter Five introduces the research design for this study, including the procedures for data gathering and analysis.

In Chapter Six, I discuss empirical studies of subjective holistic scoring and one of existing rating scales, the FCE scale. In the present chapter I have taken the theoretical literature as the basis for my argument that none of these methods of scoring is satisfactory for the context in question, and that there is a need to develop a new rating scale. However, this should be verified empirically. Therefore, it is desirable to investigate the ratings (quantitative aspects) and the rating process (qualitative aspects) when using subjective holistic scoring method and the FCE scale for assessments.

In Chapter Seven I discuss the various steps involved in developing a rating

---

[4] No doubt there are other themes to consider in addition to these three aspects before trying to develop a rating scale, for example for what purpose the scale is to be used. However, I assume that these aspects can be covered in the discussion.

scale: how I coded the scripts obtained; how I statistically analysed the codings to find the characteristics of each band; how I developed the rating scale on the basis of these features; and finally, how I revised it on the basis of feedback from the raters.

In Chapter Eight I offer empirical findings on the newly developed rating scale. Here the investigation focuses on its practicality, reliability and validity. After the validation process, I compare it with the FCE scale to explore the differences between them.

In Chapter Nine I summarise this study, discuss its limitations and make suggestions for further development and research.

### 1.8 Definition of terms

This section provides definitions of four terms that will be used throughout this thesis. The term 'to rate' means, "to exercise judgement about a performance" [performance means writing in this context] (Davies *et al.*, 1999: 160). This can be interpreted as 'to assess' or 'to judge', so in this study these three terms will be used interchangeably. A 'rater', therefore, should be understood as a person who judges or assesses writing.

As for the definition of a '(rating) scale', I follow Davies *et al.* (1999) and North (1995): Davies *et al.* (1999) define it as "a scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged. Like a test, a proficiency (rating) scale provides an operational definition of a linguistic construct such as proficiency" (p.153). North (1995) writes that 'a (rating) scale' means that the scale functions to assign "a grade in a test to which descriptions have been added for each level" (p. 65).

The term 'band' can be understood as the level which is "a measure (e.g., 1 to 9 or A to E) or description of the proficiency or ability of a test taker, normally as described on some kind of scale and determined on the basis of test performance" (Davies *et al.*, 1999: 107).

I also use the term 'subjective holistic scoring'. Hamp-Lyons (1991b) defines 'holistic scoring' as follows:

> In holistic scoring, each reader of a piece of writing reads the text rather quickly (typically one minute or less per handwritten page) and assigns the text a single

> score for its writing quality. This may be done wholly subjectively, or (and more commonly nowadays) by reference to a scoring guide or rubric, in which case it is often known as "focused holistic scoring" (p.243-4).

Given that holistic scoring can be either wholly subjective or by referring to a scoring guide, I use the term 'subjective holistic scoring' for the former, and 'holistic scoring' for the latter. In this study 'subjective holistic scoring' means that after reading a script quickly, a rater assigns it a single band according to his/her own subjective criteria, which may be based on his/her philosophy on writing ability and previous experience as a rater or teacher, without reference to any formal rating scales. It could be understood as a similar term to 'impressionistic global scoring' (Harris *et al.*, 1988). In the literature review in Part A, however, the term 'holistic scoring' could refer both to wholly subjective scoring and to the use of a reference guide, as described by Hamp-Lyons (1991b), because differentiation is not made in literature.

# CHAPTER TWO. WRITING ABILITY

## 2.1 Introduction

It is common practice among language testing researchers trying to develop a test or rating scale to start by defining the construct that they intend to tap through the test and assess using the rating scale (Hill, 1995), because the construct is generally considered to be "a broad basis for development and use of language tests and language testing research" (Bachman, 1990: 81).

In the literature the definition of writing ability – the construct in question for this study – can be seen from two perspectives, the theoretical and the pedagogical. The former will be reviewed in section 2.2, and the latter in section 2.3. This will be followed by a discussion in section 2.4 of the definition of writing ability used in the English Writing course objectives, in terms of these two perspectives.

## 2.2 The definition of writing ability from a theoretical perspective

I will begin this section by discussing a broader concept, language ability, before narrowing the discussion down to writing ability, as many relevant studies have this under concern.

Although the importance of defining language ability is recognised, it is difficult to define and measure it substantially, primarily because it is intangible and cannot be observed, and partly because it can be defined in various ways depending on the testing purpose and context.[5] In addition to these obstacles, researchers in applied linguistics have put forward various definitions of language ability (known as communicative competence since the 1970s), which has made it more difficult to define language ability in a few words.

Among the definitions suggested since the advent of the communicative approach, the most influential has been that of Canale and Swain (1980). They take a communicative approach to language ability and include the factor of 'effective use of a language for communication' in defining it. According to them, language ability, which they call communicative competence, can be divided into grammatical

[5] Horowitz (1991) contends that writing ability is task-specific and discipline-specific, and "may vary so much from task to task and from discipline to discipline" (p.72), putting forth evidence from other studies (Conklin, 1982; Glaser, 1984; Mayer, 1984) to support his argument.

competence, strategic competence and sociolinguistic competence which is further broken down into sociocultural competence and discourse competence (Canale (1983) later separated 'discourse competence' from sociolinguistic competence, creating four sub competences: grammatical competence, discourse competence, sociolinguistic competence and strategic competence). Canale and Swain recognise that language ability does not mean the ability to just recognise language, but the ability to *use* it communicatively in a given context. This tradition, with some revision, was taken up by Bachman (1990), Bachman and Palmer (1996) and Grabe and Kaplan (1996).

The definitions presented by Bachman (1990) and Bachman and Palmer (1996) are intended for general language ability. They have been very influential in language testing research. Grabe and Kaplan (1996) also follow this tradition, but they do this, focusing on writing skills. They propose a model/taxonomy of writing which includes the similar subcompetences to those in the previous definitions of communicative competence: linguistic knowledge, discourse knowledge and sociolinguistic knowledge. Linguistic knowledge covers the structural knowledge of the language in terms of written code, phonology, morphology, vocabulary and syntactic knowledge. Discourse knowledge refers to the knowledge of how writing is constructed to achieve coherence on both intra-sentential and inter-sentential levels. Sociolinguistic knowledge refers to knowledge of how language is used functionally in a variety of settings (see Appendix 1 for more details of their model).

From a theoretical viewpoint, writing ability is defined in terms of communicative competence in which 'effective use of a language for communication' is considered and which is assumed to consist of various types of knowledge or competence, that is, grammatical/linguistic knowledge, discourse knowledge and sociolinguistic knowledge.

## 2.3 The definition of writing ability from a pedagogical perspective

### 2.3.1 Classifications of approaches to the teaching of writing

Approaches to the teaching and assessment of writing will now be surveyed. These will be considered first, because the definition of writing ability from a pedagogical viewpoint depends on the approach taken.

Various approaches to the teaching of writing have developed in tandem with the development of theories of linguistics and psychology. Some of these are listed in Table 2.1 below.

Table 2.1 Pedagogical approaches to the teaching of writing

| Author | Classification of approaches |
|---|---|
| Raimes (1983) | Controlled-to-Free approach |
| | Free-writing approach |
| | Paragraph-pattern approach |
| | Grammar-syntax-organisation approach |
| | Communicative approach |
| | Process approach |
| Silva (1990) | Controlled composition approach |
| | Current-traditional rhetoric |
| | Process approach |
| | Academic-purposed writing approach |
| Johns (1990) | Process approach |
| | Interactive approach |
| | Social constructionist view |
| Tribble (1996) | Traditional text-based approach |
| | Process approach |
| | Genre approach |
| Nunan (1999) | Product-based approach |
| | Process-based approach |
| | Discourse-based approach |
| | Reader-based approach |
| Hyland (2002) | Text-oriented approach |
| | Writer-oriented approach |
| | Reader-oriented approach |

As can be seen in the table, there seem to be many approaches to the teaching of writing. They can be reduced to three main approaches: product/text-oriented, process/cognitive-oriented and reader/genre-oriented. They are sometimes reduced to only two approaches, the Product approach vs. the Process approach, as in Hedge (1998), but I will follow the position of classifying them into three approaches, since factors such as audience and social context have come to be considered important in writing, approaches involving these elements need to be included in the discussion.

Various definitions of writing ability have been made according to the three main approaches to the teaching of writing. The definitions of writing ability in each approach will be investigated in the next section.

### 2.3.2 The definition of writing ability

#### 2.3.2.1 Product/text-oriented approach

The product/text-oriented approach sees texts as either "acontextually autonomous

objects" (Hyland, 2002: 6) focusing on the surface structures of writing at sentence level, or discourse, emphasising cohesion and processability of text on the part of readers. The former corresponds to the traditional Product-based approach or Controlled composition approach, the latter to the more recent Discourse-based approach (Nunan, 1999) and Current traditional rhetoric (Silva, 1990) respectively.

The view of "Texts-as-autonomous objects" (Hyland, 2002: 6) refers to "the mechanistic view that human communication works by transferring ideas from one mind to another via language.....because meanings can be encoded in texts and recovered by anyone with the right decoding skills" (Hyland, 2002: 6-7). As a result, this approach focuses on the formal features of texts, and the goal of writing instruction is training in propositional explicitness and accuracy.

The Product-based approach (Nunan, 1999), Controlled composition approach (Silva, 1990), Controlled-to-free approach (Raimes, 1983) and Traditional text-based approach (Tribble, 1996) correspond to this "Text-as-autonomous objects" view and focus on the learners' final product, with error-free performances at sentence level being graded favourably and an emphasis placed on language form, i.e., grammar, syntax and mechanics. Although there are some researchers (e.g., Briere, 1966, cited in Silva, 1990) in this approach who argue for 'Free composition' (Raimes, 1983; Silva, 1990), it is the quality, rather than the quantity and fluency of writing that is mainly emphasised in this view. This view is inherited from structuralism and bottom-up processing theory. As a result, working on the basic notion that "the primary medium of language is oral: speech is language... speech ha[s] a priority in language teaching" (Richards & Rodgers, 1986: 49), writing is regarded as a secondary concern that functions as a reinforcement for oral habits. For psychological theory, the approach carries traits of behaviourism: learning is habit formation, in that learners are instructed to imitate, copy and transform models provided by textbooks and teachers (Nunan, 1999; Raimes, 1983; Silva, 1990; Tribble, 1996).

For teachers and researchers who subscribe to this "Text-as-autonomous objects" view, writing ability is defined as the ability to respond according to some authority's definition of the correct response to a given stimulus (Nunan, 1999). Put another way, it is "the ability to adhere to style-guide prescriptions concerning

grammar, arrangement and punctuation" (Nunan, 1999: 59) regardless of audience, purpose or context, on the basis of the assumption that a text can mean the same thing to all people only if it is written explicitly following the prescriptions (Hyland, 2002).

This view of "Text-as-autonomous objects" i.e., Product-based approach (Nunan, 1999) has been criticised since it is not reconcilable with the discourse analysis tradition that came after it. According to discourse analysts, it is discourse context, where the sentence is constructed, that determines how to arrange information in a sentence and what grammatical forms to use.

This "Texts-as-discourse" (Hyland, 2002: 10) view corresponds with the Discourse-based approach (Nunan, 1999), Paragraph-pattern approach (Raimes, 1983), and Current-traditional rhetoric (Silva, 1990), and it was introduced in the mid-1960s on the basis of the awareness that "there was more to writing than building grammatical sentences" (Silva, 1990: 13). This view is also product-based, just as the "Text-as-autonomous object" view is, in that the emphasis is on the composed *product* rather than on the process of composition. However, whereas the "Text-as-autonomous object" view emphasises the production of isolated grammatical structures, this "Text-as-discourse" view focuses on the organic relationship between discourse and grammar beyond sentence level. This "Text-as-discourse" view also stresses that learners are given samples of discourse so they can find out "how to use their knowledge of grammar in the construction of coherent texts" (Nunan, 1999: 290) and help the learners to recognise the function of sentences and paragraphs in a discourse. Accordingly, this approach intends to teach that writing is not a collection of separate sentences, but a connection of interrelated sentences producing a coherent discourse. In this vein, Raimes (1998) asserts that the central concern of this view is the logical organisation of writing. Learners are, therefore, trained to seek "to discover how writers use patterns of language options to accomplish coherent, purposeful prose" (Hyland, 2002: 10). This approach is still dominant in both writing textbooks and writing courses today. According to this "Text-as-discourse" view, writing ability is the ability to create coherent and cohesive discourses, following the prescribed patterns on how to develop and organise discourses.

To sum up, in both of these two views, writing ability is the ability to produce "acontextually" (Hyland, 2002: 6) correct forms of language, following the prescribed patterns either at sentence or discourse level.

#### 2.3.2.2 Process/cognitive-oriented approach

In the 1960s the product/text-oriented approach was criticised on the grounds that it neither fostered the writer's thought or expression, nor described adequately the composition processes (Silva, 1990). As a result, a process/cognitive-oriented approach emerged. This approach centres on what the writer does during writing. This approach, which is commonly known as the Process approach (Johns, 1990; Nunan, 1999; Raimes, 1983; Silva, 1990; Tribble, 1996) can be roughly divided into three subcategories: Expressivist, Cognitivist and Social (Situated) strands (Grabe & Kaplan, 1996; Hyland, 2002; Johns, 1990).

The first view, Expressivism, reached in its zenith in the 1960s. Teachers who subscribe to this view encourage students to develop power over their own writing without being directive, assuming that writing is a creative act and that the process is important as a discovery of the true self (Berlin, 1988). As Grabe and Kaplan (1996) note, learners are encouraged to look for their own authentic voices and to express them freely. Accordingly, the writing activities employed by those subscribing to this view are likely to be personal essays and journal writing, which are suitable for self-discovery (Johns, 1990). From this position, writing ability can be defined as the ability to express oneself freely.

After this view, the Cognitivist view that was concerned with the writing process *per se* emerged in the early 1970s, with the first language writers (Grabe & Kaplan, 1996).[6] After Emig's pioneering work (1971, 1983) on this view, many studies (e.g., Bereiter & Scardamalia, 1987; Hayes, 1996; Hayes & Flower, 1980) dealt with a cognitive model of the writing process. Among the most influential are Hayes and Flower (1980) and Bereiter and Scardamalia (1987).

Based on protocols, transcripts and videotapes of students talking aloud during writing, Hayes and Flower (1980) developed a model of the writing process

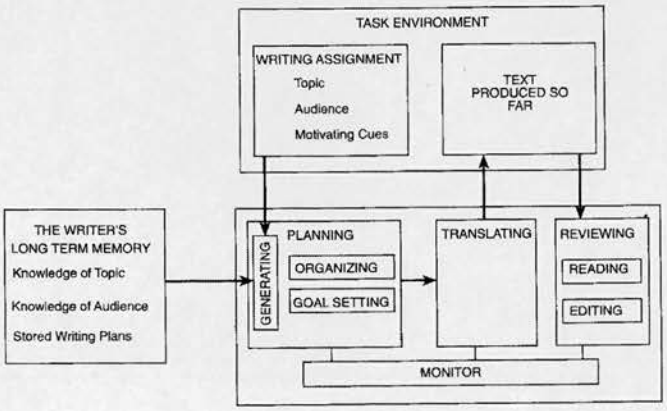
---

[6] The Expressivist view was still around in the 1980s and 1990s, and therefore existed concurrently with the Cognitivist view.

illustrated here in Figure 2.1 below. This model is composed of three parts: the composing processor, the task environment, and the writer's long-term memory. The composing processor, through which written texts are operationally generated, includes three steps, planning, translating and reviewing, all of which are individually managed by *a monitor.*

The problem with this model, however, is that "writers are not likely to be uniform with respect to their processing preferences and cognitive abilities; [...] a protocol analysis approach [which was used by Hayes and Flower] may not be a valid primary methodology for the study of the writing process to the extent that Flower and Hayes claim [...] [or at least from a more moderate perspective] it cannot be the primary source of evidence for a theory of the writing process" (Grabe & Kaplan, 1996: 92-3).

Figure 2.1 Model of writing process (Hayes & Flower, 1980: 11)

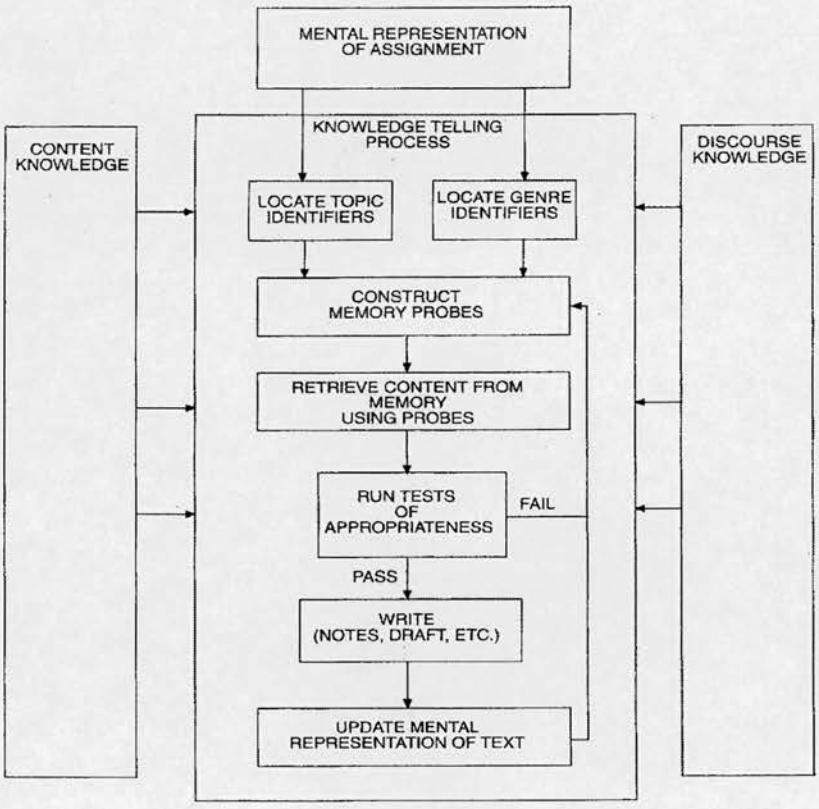



[Originally Fig. 1.5]

Pointing out the problems of Hayes and Flower's model, Bereiter and Scardamalia (1987) sought a model of the writing process that would help understand what writers actually do and why different writers write in different ways. Through their models they seek to explain what unskilled writers and skilled writers respectively do while writing. They made a distinction between 'knowledge telling' and 'knowledge transforming'. The former is a kind of writing that involves little planning and revision, which any fluent speakers of a language can carry out, even children and adolescents who are not trained to write intensively. A model of such

knowledge is shown in Figure 2.2 below.

Figure 2.2 Structure of the knowledge-telling model (Bereiter & Scardamalia, 1987: 8)

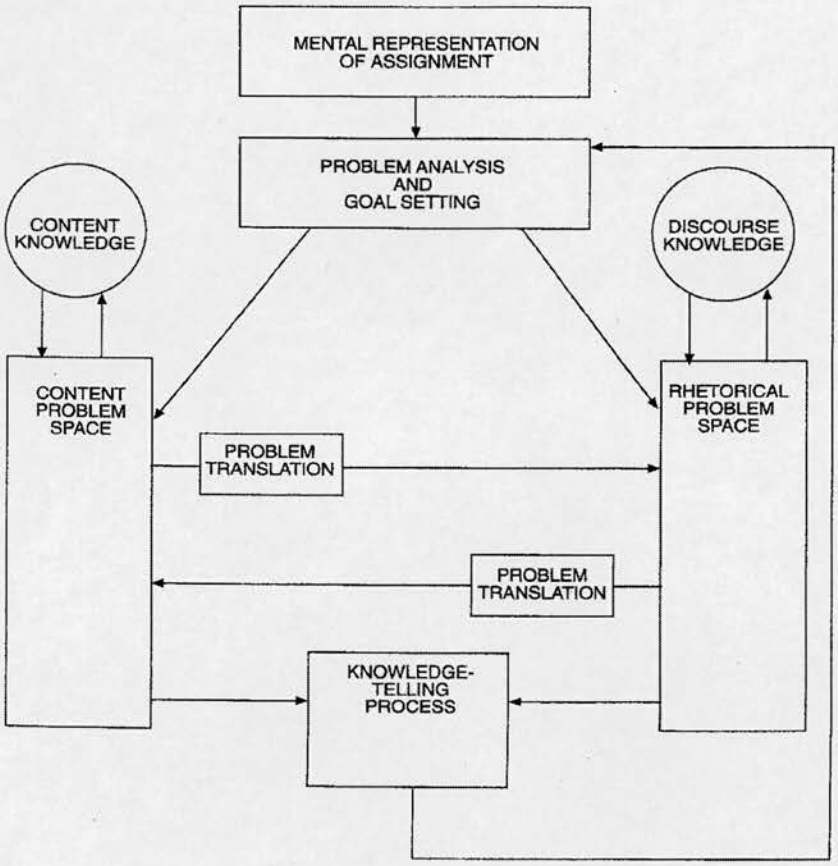



[Originally Fig. 1.1]

On the other hand, the latter, i.e., knowledge transforming, requires a great deal of effort and skill, and cannot be achieved easily. A model of such knowledge is displayed here in Figure 2.3 below.

As can be seen by comparing Figure 2.2 with Figure 2.3, the latter can be said to be an extended version of the former, as the latter includes knowledge telling plus other elements. The difference between them, consequently, lies in the added part: namely, whether there are problem analysis, goal setting and problem translation in the model. This meta-cognitive elements lead to problem solving activities in two subsequent domains, that is, the content problem space and the rhetorical problem space, both of which interact with each other in a two-way attempt to find solutions to the problems of either content or discourse.

Figure 2.3 Structure of the knowledge-transforming model (Bereiter & Scardamalia, 1987: 12)




[Originally Fig. 1.2]

The investigation from this Cognitivist view is ongoing, although it now uses more developed methods than traditional ones such as the think-aloud method. For example, Glendinning and Howard (2001) explored the actual process of writing (with L2 learners), using Lotus ScreenCam,[7] inspired by Hairston (1982):

> We cannot teach students to write by looking only at what they have written. We must also understand how that product came into being, and why it assumed the form it did. We have to try to understand what goes on during the act of writing… if we want to affect its outcome. We have to do the hard thing, examine the intangible process, rather than the easy thing, evaluate the tangible product (Hairston, 1982: 84).

[7] "Lotus ScreenCam turns your PC into a VCR that records every click, scroll and action on your screen"[(http://www.lotus.com/products/screencam.nsf/)]. The result is a "movie" which can be played back like a video tape. In the case of word-processing, every word typed, deleted, cut and pasted, or dragged to another location is recorded. Furthermore, a "sound-track" can be added if the PC has a microphone and sound card (Glendinning & Howard, 2001: 43).

The Cognitivist views discussed so far are laudable in that they explored the "intangible" (Hairston, 1982: 84) writing process. However, they do have certain shortcomings. First, they were developed with first-language writers so that the issue of L2 learners was not dealt with at the time (recent studies, such as Glendinning and Howard (2001), include studies on L2 learners). Another criticism was that they paid little attention to the social contexts that help specify the particular writing purpose. As a result, the third view within the process-oriented approach emerged in the 1980s: the Social (situated) view.

This view seeks to investigate the writing process on the basis of the assumption that writing is a situated act. Hyland (2002) explains it as follows:

> Research here seeks to move beyond the possible workings of writers' minds and into the physical and experiential contexts in which writing occurs. This view rejects the myth of the isolated creator and sets out to describe how 'context cues cognition' (Flower, 1989). Of crucial importance is the emphasis placed on a notion of context as the 'situation of expressions' (Nystrand, 1987). [..] The goal is to describe the influence of this context on the ways writers represent their purposes in the kind of writing that is produced (p. 30-1).

Since researchers advocating this orientation mean to observe what is *actually* occurring, without imposing an *a priori* framework according to which observations are illustrated, they usually use ethnographic research methods for their studies.

Of these three views (Expressivist, Cognitivist and Social (situated)), the Cognitivist perspective has occupied a dominant position. Johns (1990) believes that its influence on modern ESL classrooms cannot be exaggerated. Taking this approach, researchers and teachers aim to help learners to develop writing skills *per se*,[8] rather than help them produce correct writing, by making the writing process of skilled writers explicit through research (e.g., Bereiter & Scardamalia, 1987) and replicable by unskilled writers.

Writing ability in this process/cognitive-oriented approach is defined as the ability to initiate and evolve ideas and then use certain revising and editing practices to develop them to maturity in a given context.

---

[8] Some researchers, therefore, contend that the cognitivist tradition of the teaching of writing is not appropriate for many adult L2 learners because they have such a fully developed knowledge of the writing process in their first language that they do not have to learn how to write; instead they need to learn about the conventions and constraints of the target language (Tribble, 1996).

### 2.3.2.3 Reader/genre-oriented approach

In this reader/genre-oriented approach, the elements of audience and social context are added to the teaching of writing.[9] In this approach a writer who recognises context and audience (i.e., the discourse community) for which and for whom the written product is produced, is likely to appreciate the importance of rhetorical knowledge such as format, style and content in matching a text to a social purpose and shaping a successful text. This emphasis on the constraints of form and content is related to the notion of 'genre' (Tribble, 1996).

The term 'genre' was originally devised to classify literary works into poems, novels, dramas, etc., according to a particular style, form or content. It has recently gained currency outside literature, especially in film, cultural media, music and other fields. The term has come to be used for the classification of types of common spoken and written discourse. In written discourse, it has been defined as "a discourse type that [has] identifiable formal properties, identifiable purposes and complete structure (i.e., a beginning, a middle and an end)" (Grabe & Kaplan, 1996: 206), "classes of communicative events consisting of texts, text-roles and the environments wherein those texts are produced, interacted with and received" (Swales & Horowitz, 1988, cited in Horowitz, 1991: 73), or "a communicative act which is culturally conditioned and institutionalised in its form, which can be recognised or classified from its communicative purpose" (Fulcher, 1997: 91). Gee (1997) quotes the definition given by Martin *et al.* (1987) as follows:

> [It means] a staged, goal oriented social process. Most members of a given culture would participate in some dozen of these. [....] Genres are referred to as *social processes* because members of a culture interact with each other to achieve them; as *goal oriented* because they have evolved to get things done; and as *staged* because it usually takes more than one step for participants to achieve their goals (p. 59).

Swales' (1990) proposition in his book *Genre Analysis* is also helpful in allowing a practical grasp of the term for writing instruction.

---

[9] This does not mean that the notion of audience was only taken into consideration for the first time with the emergence of the reader/genre-oriented approach. All the approaches mentioned above have perceived the role of audience in each approach (see Johns, 1990 for details), but it is given most attention in this approach.

> *A genre comprises a class of communicative events, the members of which share some set of communicative purposes*. These purposes are recognised by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. Communicative purpose is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rhetorical action. *In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience* (p. 58) [emphasis added].

Given Swales' definition of genre as a communicative event in accordance with purpose and structure, style, content and intended audience, examples of genres in written discourse include fiction novels, grant applications, progress reports, course syllabi and survey articles and so on (Grabe & Kaplan, 1996).

Based on this notion of genre, the Genre-based approach to the teaching of writing (e.g., Gee, 1997; Hyland, 2004; Johns, 2002) has arisen from the concern as to whether the process/cognitive-oriented approach, which was sweeping through the writing curriculum, fully addresses the needs of learners who need to write an effective text that meets the readers' specific expectations of form, content and style (Tribble, 1996).

The Genre-based approach is "based on satisfying the demands of the discourse community" (Weir, 1993: 130), and is consequently "based on a selection of relevant genres [...] which entail[s] a departure from an exclusive preoccupation with narrative/expressive writing and give[s] recognition to factual writing" (Gee, 1997: 25).

However, this approach has been criticised as too prescriptive, as taking a normative approach to the production of texts and focusing on the final product like the product/text-based approach[10] even though as Fulcher (1996d) claims, "writing is process *and* product" (p. 46) [emphasis added]. Although the Genre-based approach has been criticised at times, it has held sway with writing teachers and

---

[10] However, this is not an insurmountable problem. First, as Nunan (1999) proposes, the Genre-based approach is an issue of syllabus design, not classroom action, whereas the Process-based approach is a methodological issue concerned with classroom action. Therefore, classroom action can be organised according to the Process-based approach, whilst syllabus design can be organised according to the Genre-based approach. In that case, classroom action would not be prescriptive. Second, Tribble (1996) argues that the genre itself is not static but dynamic, so social practices regarding genre and texts feature fluidity. As a result, this fluidity and awareness on the part of teachers may encourage them to use a different teaching style from the one for the product/text-oriented approach. Accordingly, it cannot be always prescriptive.

researchers since it was first developed in Australia in the 1970s. Weir (1993) also mentions that there is a strong move towards the use of the Genre-based approach in courses on writing in English for Academic Purposes (EAP).

This Genre-based approach, which emphasises awareness of the reader, regards the successful writer as one who is able to reasonably assume what the reader knows and expects, to seek a balance between his/her writing purpose and the reader's expectations, and to satisfy the reader's rhetorical demands.

Thus, according to this approach, writing ability is defined as the ability to perform writing tasks for a given purpose, which satisfy a given discourse community in terms of structure and content of discourse and communicate functionally.

To sum up, I have discussed writing ability in terms of two perspectives, the theoretical and the pedagogical, and shown that there is not just one definition of writing ability that is commonly accepted by teachers and/or researchers. Rather, its definition depends on both linguistic theory and approaches to the teaching of writing.[11] In the next section I will investigate the definition of writing ability used in the English Writing course objectives in terms of these two perspectives.

## 2.4 The definition of writing ability used in English Writing course objectives

In this section, I will try to find which model is closest to the concept of writing ability implicit in the course. As mentioned in section 1.6.3, the writing ability implicit in the course is the ability to express one's thoughts and feelings in an organised, accurate and fluent manner in various contexts for effective communication. From the theoretical perspective, this writing ability is in accordance with the concepts of communicative competence, especially compatible with the model of Grabe and Kaplan (1996), in that the phrases "in various contexts in an organised manner", "accurately" and "fluently" (Lee *et al.*, 2001: 195) are respectively compatible with discourse knowledge, grammatical knowledge and

[11] I believe that the purpose of the learning of writing is one of several other factors that could affect writing ability.

sociolinguistic knowledge in their model.

Next, from the pedagogical perspective, it appears that the definition of writing ability implicit in the course is closest to the product/text-oriented and reader/genre-oriented approaches, in that the course focuses on both sentence structure and discourse – as revealed in the stress on writing "in an organised manner and accurately" (Lee *et al.*, 2001: 195), and on communication with readers – as revealed in the stress on "effective communication" in various genres and contexts.

In sum, in light of these two perspectives, the definition of writing ability implicit in the course can be expressed as the ability to produce both "acontextually" (Hyland, 2002: 6) and contextually correct forms of language following prescribed patterns both at sentence level and at discourse level, so as to communicate with readers functionally.

## 2.5 Summary and conclusions

In this chapter I have reviewed the definitions of writing ability in the literature and tried to investigate how the definition of writing ability implicit in the English Writing course in Korea could be explained in the light of previous research. To begin with, I discussed the definitions in previous research from two perspectives, the theoretical and the pedagogical. Of the former, I considered the views of Canale and Swain (1980), Canale (1983), Bachman (1990), Bachman and Palmer (1996) and Grabe and Kaplan (1996) (apart from Grabe and Kaplan, these authors discuss linguistic ability, which is a broader concept than writing ability). For the pedagogical perspectives, definitions of writing ability varied according to three main approaches to the teaching and assessing of writing: product/text-oriented, process/cognitive-oriented and reader/genre-oriented approaches.

I then considered the writing ability implicit in the course in terms of these two perspectives. From the theoretical perspective I can conclude that it most closely corresponds to Grabe and Kaplan's (1996) ideas, in that it can be expressed as being composed of grammatical knowledge, discourse knowledge and sociolinguistic knowledge.

From a pedagogical perspective, I found that it was based on product/text-oriented and reader/genre-oriented approaches, and that it can be explained as the

ability to produce both "acontextually" (Hyland, 2002: 6) and contextually correct forms of language following the patterns prescribed at both sentence and discourse level, so as to communicate with readers functionally in various genres and contexts.

# CHAPTER THREE. WRITING ASSESSMENT

## 3.1 Introduction

As mentioned in Chapter One, the guidelines to the English Writing course at FLHSK make it explicit that the course aims to help students develop writing ability through "free writing". In order to do assessment towards this goal, the test should take the form of free and extended writing rather than guided or controlled writing or a multiple-choice type of writing test. Before exploration of rating scales, we need to discuss this writing test to help understand the context in which the new rating scale will be used.

In section 3.2, I will review the history of writing assessment. In section 3.3, I will narrow down the discussion to direct tests, which are employed in many schools, universities and colleges, as well as for the English Writing course in question. Specifically, I will discuss the issues with direct tests: reliability vs. validity, and factors affecting the validity of writing tests. In section 3.4, portfolio assessments will be examined, and in the final section, section 3.5, I will offer some conclusions.

## 3.2 The history of writing assessment

Writing assessment *per se* in L1 is not a new discipline. It actually dates back thousands of years, although writing assessment then was "assessment *through* writing" rather than "assessment *of* writing" (Hamp-Lyons, 2002: 6). According to Hamp-Lyons' review (1990, 1991a, 2002) of the history of writing assessment, writing assessment in the form of "assessment *through* writing" was carried out both in Asia and Europe until the Middle Ages, as a means of either selecting servants at the Imperial Court or controlling the teaching in Catholic schools. But it was only around the 1900s that writing assessment in the form of "assessment *of* writing" arose for the first time. The assessment then was in the form of direct tests. Over the next 30 to 40 years test-takers were asked to actively write something.

With the advent of the structuralist-psychometric era in the US in the 1950s and 1960s, writing tests fell from favour, particularly direct writing tests. Hamp-Lyons (1990) points out that this was due to the scoring procedure rather than the testing method itself. She explains this as follows: there are three phases common to all tests

of writing: the construction of questions; answering questions; and the scoring process. The first two of these three processes are necessarily subjective (therefore, Hamp-Lyons (1991a) points out that it is wrong to describe indirect tests as 'objective tests', because the design process of indirect tests is as subjective as that of direct test). The last process can be objective in indirect tests, which simply ask test-takers to recognise a correct answer when given several choices, but cannot be objective in direct tests, which involve subjective judgements. As a result, during the psychometric era the use of direct tests was replaced by indirect tests, which seemed to fit the prevailing belief in 'objectivity'.

Hamp-Lyons (1991a) characterises indirect tests as tests "where there is no room for personal interpretation by the test-taker since possible answers are provided and the 'correct' one already decided upon" (p. 6). A typical example of indirect tests for writing assessment is the written expression part of TOEFL in the past, which takes the form of a multiple-choice test. Test-takers were given two types of task: (1) to choose the most appropriate of several words/phrases to fill the blank in a given sentence; (2) to recognise a part of a given sentence that was written ungrammatically. They were never invited to actively write something. As White (1994) notes, this kind of choosing activity is entirely different from that of writing in the real world. To enter the world of multiple-choice tests means to accept a worldview in which questions have one correct answer to be selected from given options. When we write, however, we do something totally different. Hyland (2003) explains the difference between the activities involved in direct tests and indirect tests as follows:

> When we write, we inhabit a quite different world, one in which we must generate and select from many options ourselves and in which most answers are at best partially true. However simple the writing task is, we must select appropriate vocabulary, frame sentences, connect ideas and express our own views. As writers, we know that life is complex and that simple answers are usually wrong. But if we use this perspective of the writer on a multiple-choice test, we are likely to get into trouble; we may see that, under some circumstances, the wrong answers could be right and the right answers wrong. But if we are good test takers, we sink those perceptions, along with all the other ambiguities and problems of life and focus only on the single question that matters in the test world (p. 174).

Indirect tests achieved high reliability and became common as standardised tests

for L2 writing assessment in the US. In the UK, however, indirect writing tests were widely opposed, and direct tests were retained despite the fact that they are more complicated to conduct (Hamp-Lyons, 2003). Ways of raising the level of reliability in scoring, which was considered one of their defects, were pursued.

Since then, as the level of reliability rose and emphasis was increasingly put on the communicative element of language in the 1970s, direct tests have been revived in L2 writing assessment in the US. This was initiated by the development of the English Language Testing Service (ELTS) in the UK in 1980, and the subsequent production of a rating scale (Testing ESL Composition by Jacobs *et al.*, 1981) for direct writing tests in the US. The trend towards direct tests was continued by the TEEP (Weir, 1990) and the introduction of the TWE as an optional test in TOEFL in 1986. Thus, Calson and Bridgeman (1986) review, since the end of the 1980s the pendulum has clearly swung from indirect tests back to direct tests, which approximate real discourse more closely than indirect tests, and permit the evaluation of writing skills such as organisation, coherence and the elaboration of ideas, which were not measured satisfactorily by indirect tests.

These two types of tests are obviously opposite in nature. But this does not mean that either of them is always either good or bad, nor that their merits are always guaranteed. Conlan (1986) notes that direct tests have face validity, which gives them an advantage over indirect tests, which do not have it. Therefore, direct writing assessments have been generally accepted as valid methods (Schoonen *et al.*, 1997). However, even direct tests can sometimes be invalid. For example, in programmes assessing basic skills at the ninth-grade level in the US, learners were asked to write a letter of application in response to an advertisement printed in the test booklet. For a while it was fashionable to include this kind of assignment in the US, but while it might look like a practical writing task, something done in the "real" world, writing a letter of job application is not something that thirteen or fourteen-year-olds normally do, so this test cannot be seen as a valid measure of their writing ability.

As for indirect tests, on the other hand, they are often described as "efficient", in that a great many questions can be presented to students and answered in the limited testing time available. Thus, they may achieve efficiency that is its strength. However, this efficiency can be valid only if the tests have construct validity. They may lack

construct validity as well as face validity, if they contain questions that have not been taught on the course. In this case, the loss (of both face validity and construct validity) is greater than the gain (in efficiency).

Thus, neither direct tests nor indirect tests are good in their own right. Whether they are good or not depends on their application and situation. Conlan (1986) describes them as partners, and contends that test practitioners should either choose one or combine both of them depending on the purposes of the assessment, the characteristics of the test-takers, the budget and resources available and so on. Because it may be useful to be aware of the characteristics of these two test types in order to take advantage of them appropriately, Conlan (1986) summarises their characteristics as shown in a table below, reported here as Table 3.1.

Table 3.1 The characteristics of direct tests and indirect tests for writing ability assessment[12] (quoted from Conlan, 1986: 117-8)

| | Direct tests | Indirect tests |
|---|---|---|
| Method of measurement | Direct---candidate asked to perform task to be measured | Indirect---measurement relies on correlation between test performance and actual task performance. |
| Skills measured | Unlimited---candidate must compose, organise, marshal evidence, spell, punctuate, etc.<br><br>Total---all aspects of writing can be measured simultaneously. Permits use of complete essay. | Limited---certain aspects of writing cannot be measured (e.g., ability to marshal evidence, ability to set proper tone).<br>Fractionated---writing skill must be separated into parts to be measured independently [13]. Relies mostly on aspects of writing that can be measured in the sentence. |
| Time required | Depends on kind of writing required; not less than 20 minutes per question. | Depends on item types used; can require as little as 30 seconds per item. |
| Sampling done | Limited by time---no more than 3 samples per hour; best to have fewer. Candidate who misinterprets or does not understand question misses major part of test. | As many as 100 items per hour. Candidate who misses one question is not in serious jeopardy. |

[12] The original title is "essay and multiple-choice questions in tests of writing ability". Since "essay" and "multiple-choice questions" are representative examples of direct tests and indirect tests respectively, I have adapted the title, for the flow of discussion.

[13] This may allow indirect tests to provide diagnostic information on a test-taker's writing. It is, however, worth noting that the diagnostic information from the indirect tests relates to how to "recognise" aspects of writing at sentence level, whilst the diagnostic information from direct tests, particularly the information produced by using an analytic scale, relates to how to "actually write", i.e. the decisions made when the test-taker was writing, in terms of how to construct and organise every single sentence, which words or phrases to choose, what to write about for a given topic and so on. Therefore, given that there is a gap between "just recognition", i.e. deciding whether the use of a word or grammatical feature is correct or not, and "actual production", i.e. whether the test-taker is able to retrieve and use a word or grammatical feature, with appropriate planning or organisation, the information that could be derived from either indirect tests or direct tests is not qualitatively equal.

| | | |
|---|---|---|
| Method of scoring | Individually, by trained readers. | Can be machine scored. |
| Validity | Increases face validity by providing direct measure. By requiring actual task, extends what can be measured; thus increases validity.<br>Because sampling is limited, validity of essay used alone is less than that of essay used with multiple-choice. | High correlation between scores on multiple-choice and essay tests. |
| Reliability of scoring | Reliance on subjective impression reduces reliability. | Same as other machine-scored tests. |
| Test reliability | Limited by scoring reliability; length of test | Can be above .90; a one-hour test can be 100 items long. |
| Cost | Increases for scoring (housing and paying readers, etc.) and special procedures (new answer sheet, new systems design, etc.). | Same as regular machine scoring. |
| Time for scoring | Readers can read 20-minute essays at rate of 38 per hour; reading day about 6 hours. | Same as regular machine scoring. |
| Reaction of English faculty | Approval | Hostility and distrust. Many believe (1) multiple-choice tests are so limited in what they measure that they reduce writing to level of subjective-verb agreement, (2) tests are exercises in error-hunting, and (3) the way to measure writing is to have people write. |
| Influence on curriculum | Thought to encourage requirement of actual writing in schools | Thought to encourage exercises in error detection as method of teaching writing rather than encouraging writing of compositions. |

[Originally No Table No.]

Understanding the characteristics of these tests may help them to be used appropriately. As Conlan (1986: 117-8) illustrates, for example, in cases when the writing skills of large numbers of students are to be tested and the budget for scoring in direct tests is limited, indirect tests could be better as screening devices to separate out those who will not need to have their scores further examined by a direct test.

However, as concern for validity is now dominant (Connor-Linton, 1995), it is true that direct tests, which potentially have face validity in many situations, are favoured amongst test developers over indirect tests for writing ability assessment. Therefore, the next section onwards will focus on direct tests.

Direct tests include timed impromptu tests and alternative tests such as portfolio assessments. In the following sections direct tests will specifically mean timed impromptu writing tests that require test-takers to write a single piece of writing within a timed test situation (Weigle, 2002). I will discuss them in more detail in terms of the main issues they raise: reliability vs. validity, and the factors affecting

validity.

Alternative assessment such as portfolio assessment, that has emerged since timed impromptu writing tests were criticised for being neither conducive to the curriculum nor reflecting the nature of real-world writing, which includes multi-drafting and revising (White, 1995), will be discussed separately (in section 3.4). This is because there are more differences than similarities between portfolio assessments and direct tests, in that portfolio assessments are conducted in an unsupervised situation with several pieces of writing, while the timed impromptu tests use a 'snapshot approach' with one piece of writing. Even so, portfolio assessments are closer to direct tests than direct tests are to indirect tests.

## 3.3 Issues in direct writing tests

It is obvious that direct tests are more complicated to conduct than indirect tests, especially in terms of the scoring procedure. Whereas indirect tests can be scored by machine very reliably, direct tests need to be scored by humans. As a result, the scoring in direct tests is not as reliable as the scoring in indirect tests, and although direct tests are assumed to be more valid than indirect tests, as the validity of direct tests can be affected by various factors it cannot be said that the validity of direct tests is guaranteed. These issues will be discussed further in the following sections.

### 3.3.1 Reliability vs. validity

#### 3.3.1.1 Shifts in focus between reliability and validity

Since the 1970s, when writing assessment once again took the form of direct tests (Huot, 1990b) rather than the indirect tests favoured in the US in the 1950s and 60s, the centre of interest for both testing practitioners and language test theorists has been the issue of reliability, which means consistency (White, 1995). According to Hyland (2003), reliability can be divided into two types, performance reliability and scoring reliability. Performance reliability has to do with whether the same student performs consistently on different occasions, and therefore requires multiple writing samples covering different topics and genres. Scoring reliability, on the other hand, concerns consistency among different rating occasions in rating the same piece of writing. There are two types of scoring reliability: inter-rater reliability – whether all

raters agree on the rating of the same performance, and intra-rater reliability – whether each rater rates the same performance in the same way on different occasions.

Achieving a high level of scoring reliability is not easy, because raters "engage in complex problem-solving activity whereby they 'construct' the particular scoring decision as relevant to the certain criteria and then invoke the criteria in different ways" (Torrance, 1998: 33). Therefore, many assessment criteria have been developed to raise levels of scoring reliability (e.g., Alderson, 1991; Jacobs *et al.* 1981; Weir, 1990), and rater training has been undertaken to help raters to apply criteria in the same manner (e.g., Brown, 1995; Weigle, 1994, 1998). Many studies have been undertaken to report the degree of reliability of rating in writing assessments, and to determine the effects of scoring schemes and rater training. In some cases test designers achieved an agreement between two raters of 75% or more. Therefore, it is widely agreed that direct tests are no less reliable than indirect ones, whose greatest virtue is objective and reliable scoring (Hyland, 2003).

However, it has been increasingly recognised that the issue of reliability is not the only domain deserving attention. Reliability is certainly a necessary quality in language tests, but it is not sufficient (Henning, 1987; Moss, 1994). In writing assessment the issue of validity is now being considered: whether qualities that the test is intended to measure are actually measured (Lado, 1961). This means that validity has received more attention than ever, but does not mean that concerns about reliability have been dispelled. Furthermore, given that reliability is currently not considered to be a separate concept from validity, but is seen as part of validity (see section 3.3.1.2 for further discussion), reliability still needs to be considered as evidence of validity. Therefore, while the increased interest in and concern about validity may have reduced the previous exclusive focus on reliability, they have not dispelled the need for reliability *per se*.

Apart from the discussion about the status of reliability compared with that of validity, this interest in validity has caused direct tests to be favoured over indirect tests, since direct tests are considered to be a valid way to gather information on a learner's writing ability (Schoonen *et al.*, 1997).

The discussion about reliability and validity has not been limited to writing tests,

but has covered the whole field of language testing, in which validity is the central concern. This will be further discussed in the next section.

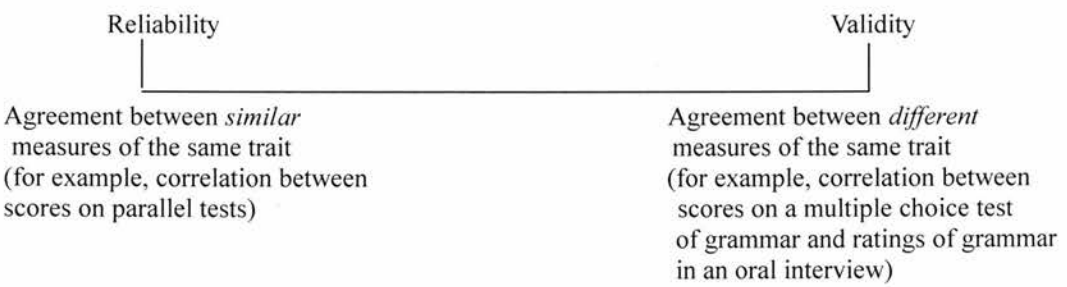
### 3.3.1.2 Further discussion of reliability and validity

Validity and reliability are seen as two key issues in language testing. In the past, reliability was considered to be distinct from, but a requirement for, validity, with each investigating a different concern. Bachman (1990) writes that:

> the investigation of reliability is concerned with answering the question, 'How much variance in test scores is due to measurement error?'[....] Validity, on the other hand, is concerned with identifying the factors that produce the reliable variance in test scores. That is, validation addresses the questions, 'What specific abilities account for the reliable variance in test scores?' Thus, we might say that reliability is concerned with determining how much of the variance in test scores is reliable variance, whilst validity is concerned with determining what abilities contribute to this reliable variance (p. 239).

Bachman (1990) illustrates this distinction between reliability and validity as shown in a figure below, reported here as Figure 3.1.

Figure 3.1 Relationship between reliability and validity (quoted from Bachman, 1990: 240)




[Originally Fig. 7.1]

However, as Bachman (1990) points out, the distinction between reliability and validity is not considered so clear now, firstly, because the distinction between the test methods is not entirely obvious, and secondly, because the distinction between the abilities we intend to measure and the facets of the measurement procedure is not very clear, either. For example, in the former it is unclear whether the correlation between concurrent scores on two cloze tests based on different passages is an issue

of reliability or validity. With the latter, in the case of oral interview tests where the definition of communicative language ability includes the productive modality and oral channel, since these trait and method factors are bound together, it is not clear whether the test score is to be interpreted in terms of reliability or validity. Therefore, Bachman (1990) concludes that reliability is not clearly distinguishable from validity.

In addition to this, the assumed relationship between reliability and validity has changed. In the past, reliability was seen as a requirement for validity, so it was considered that while reliability of a test did not guarantee its validity, it could not be valid unless it was reliable. However, reliability is now considered as one type of evidence for validation. Citing Lado (1961), Davies and Elder (2005) also mention that while in the past, reliability was necessary but not sufficient, and that sufficiency depended on validity, the current fashion is to see reliability as a part of validity.

Therefore, it appears that reliability and validity should not be distinguished as separate and complementary concepts. Given this, I will discuss reliability in the realm of validity and validation,[14] rather than make it separate from the scope of validity. That is, since reliability (that means scoring reliability (Hyland, 2003) here) means coincidence between more than two ratings of a performance, it can correspond to concurrent-related evidence of validity.

In this context, validity is a central concept for measurement, and has been discussed as such in the field of educational measurement as well as in the field of language testing. The discussion of validity generally refers to its definition and validation methods. The definition has changed since Lado (1961). According to Chapelle's review (1999) of the history of the concept of validity, in the early era Lado (1961) defined it as the issue of "Does a test measure what it is supposed to measure? If it does, it is valid" (p. 321). It was considered that there were three kinds of validity: criterion-related validity, content-related validity and construct validity. At the time, correlation methods were considered as central for validation.

This kind of definition persisted through the 1980s in the field of language testing, but a variety of methods were employed for validation, such as gathering qualitative data (i.e., think-aloud protocol) on test-taking strategies (Cohen, 1984)

---

[14] Chapelle (1998) writes "sufficient justification of the interpretations made from test performance in an operational setting.[...] is validation" (p. 49).

and comparing test methods (Shohamy, 1984). Additional kinds of validity were also proposed. For example, Henning (1987) suggested five types of validity: predictive, concurrent, content-related, construct and response. In the meantime, there were significant developments with regard to validity in the field of educational measurement. It was suggested in Messick (1989) that construct validity was a unified overarching validity rather than one of the three kinds of validity previously posited.

As a result of these developments, the 1990s saw an explicit discussion of validity in the language testing field as well. Bachman's (1990) book of this decade emphasised three aspects: that validity has to do with the inference made from the test score rather than test itself; that construct validity is an overarching concept, with content and criterion-related investigation being evidence for construct validation; and that the consequences of test use are part of the issue of validity. Chapelle (1999) follows Messick's (1989) framework in summarising the changes in the concept of validity and validation methods in the 1990s, which are shown in a table below, reported here as Table 3.2.

Table 3.2 Summary of the contrasts between past and current conceptions of validation (quoted from Chapelle, 1999: 258)

| Past | Current |
|---|---|
| Validity was considered a *characteristic of a test*: the extent to which a test measures what it is supposed to measure | Validity is considered an *argument* concerning test interpretation and use: the extent to which test interpretations and uses can be justified. |
| Reliability was seen as distinct from and a necessary *condition for validity* | Reliability can be seen as *one type of validity evidence*. |
| Validity was often established through *correlations* of a test with other tests. | Validity is argued on the basis of a number of types of *rationales and evidence*, including the consequences of testing. |
| Construct validity was seen as one of *three types of validity* (the three validities were content, criterion-related, and construct). | Validity is a *unitary concept* in which construct validity is central (content and criterion-related evidence can be used as evidence about construct validity). |
| Establishing validity was considered within the purview of *testing researchers* responsible for developing large-scale, high-stakes tests. | Justifying the validity of test use is the responsibility of *all test users*. |

[Originally Table 1]

The table above shows that while in the past validity was defined as a characteristic of a test, it is currently defined as "the degree to which our conclusions, or inferences are true" rather than the instruments and procedures themselves, shifting the focus to "the veracity of conclusions, or inferences" (Lynch, 2003: 149).

This follows Messick's definition (1988) of validity as: "an overall evaluative judgement, founded on empirical evidence and theoretical rationales, of the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores" (p. 33). Additionally, it was said that there was a set of types of validity, such as face validity, predictive validity, concurrent validity, content validity and construct validity.[15] Since it was considered that "construct validity is not an easy idea to work with", and that "in practice, there may be little difference between construct and content validity" (Underhill, 1987: 106), there were many cases where content validity was employed, especially in EAP tests, for example.[16] Nowadays, construct validity tends to be deemed as a central unitary validity,[17] and the validity types applied in the past and even reliability are used as evidence for construct validation.

Even so, there is still a difficulty with research on validity, due to the fact that this is an abstract concept. Therefore, before one says that the inference of the score from a test or a rating scale is valid or not, the concept needs to be operationalised. This is done through some evidence of validity. As Davies and Elder (2005) say in their review of Messick (1988), "excessive reliance on only one kind of validity evidence is unlikely to be seriously endorsed for long" (Messick, 1988: 35). It is, therefore, a current fashion to talk about multiple sources of evidence for validity, which Davies and Elder (2005) identify as the issue of "the unitary and the divisible" (p. 797). Chapelle (1999) summarises that Messick (1989) suggested six types of evidence. First, content analysis can be taken advantage of, which consists of experts' judgements of what a test appears to measure. Here, the evidence is supposed to demonstrate that a test is relevant to and covers a given area of content or ability. This validation is done in terms of two aspects: content relevance and content coverage. However, as Bachman (1990) points out, the problem with this evidence is that it is difficult clearly and unambiguously to identify a domain that helps demonstrate either content relevance or content coverage. Additionally, the critical limitation of this evidence is that neither content relevance nor content

---

[15] Angoff (1988) reviews the history of the emergence and development of each type of validity.

[16] On the grounds that construct validity should be used for the validation of test design and evaluation, rather than content validity, Fulcher (1999) emphasises the importance of construct validity, following Messick's (1989) framework on validity.

[17] Messick (1989) emphasises that validity is a unitary concept, and construct validity is a central overarching validity.

coverage accounts for how test-takers actually perform in the test. Put another way, it is just a *test* characteristic rather than a test score characteristic. Furthermore, even if a test is justified as valid with the evidence of content analysis, if the score is used for purposes other than an appropriate specific purpose, the score is not valid. In summary, since the evidence for validity on the basis of content analysis focuses on tests rather than test scores, it cannot be a sufficient piece of evidence for validity, even though it is necessary.

Second, the evidence from differences in test performance can be used, looking at the differences between different groups of test-takers, time, instruction and so on. In the past this evidence from different groups of test-takers was categorised as concurrent validity; now it is known as concurrent criterion-related evidence. Typical groups for this purpose are native speakers and non-native speakers of the language. Regarding the problems with this tradition, Bachman (1990) states that:

> [F]irst, we must carefully examine the basis on which we assume one group to be more proficient than another [...] there are serious problems in determining what kind of language use to consider as the 'native speaker' norm, while the question of what constitutes a native speaker, or whether we can even speak of individuals who are native speakers, is the subject of much debate. Furthermore, there is growing evidence that native speakers perform *neither* uniformly well on tests of all aspects of language ability, *nor* uniformly better than do non-natives (for example, Allen *et al.*, 1983; Bachman, 1985). Second, we must not assume that since individuals in one group are at a higher level of language ability *in general*, they will therefore be at a higher level on the specific ability in which we are interested (p.248-9).

As for differences in instruction, randomisation is key to the evidence drawn from them. That is, individuals are randomly selected from a population and randomly assigned to two or more groups, each of which is treated differently (e.g., receiving a specific instruction or no instruction). When the group that has been given an instruction regarding the specific construct shows test higher scores, the interpretation of the test is said to have construct validity.

The evidence from differences in time is known as predictive utility. This requires the collection of data to demonstrate a relationship between scores on the test and job/course performance. However, Bachman (1990) writes that it is problematic because criterion behaviour is complex and depends on various other factors as well as language abilities.

Third, empirical item or task analysis is another validation method, which involves quantitatively or qualitatively investigating the extent to which the relevant factors affect the item difficulty and discrimination.[18] Quantitative analysis is usually based on correlation evidence. Depending on whether a factor affects test-taking, the correlational evidence shows either consistent correlation or non-consistent correlation between the two aspects. Qualitative analysis methods, on the other hand, are the means of investigating the processes of test taking, and can compensate for the limitations of validation methods that are based on evidence from correlational evidence. Cohen's study (1984), which examined test-takers' test taking strategies and reactions to different items and test types using verbal self-reporting methods, is a good example for this.

Fourth, dimensionality analysis can be used. This method assesses the extent to which the observed dimensionality of response data coincides with the hypothesised dimensionality of a construct, by using methods such as the Item Response Theory (IRT, Henning, 1987; McNamara, 1996) in order to investigate the internal structure of the test. The dimensionality of a test or a hypothesised construct can be either unidimensional or multidimensional. When the dimensionality of a test and the hypothesised dimensionality of a construct are coincident with each other, the test can be said to have construct validity. Since many hypothesised constructs tend to be multidimensional, a test for these constructs needs to be multidimensional as well.

Fifth, the relationship of test scores with other tests and types of behaviour can be used as an evidence of validation. This is usually done by Multitrait Multimethod (MTMM, Campbell & Fiske, 1959), which is analysis of divergent and convergent correlations between either test methods or constructs. Convergence can be understood as concurrent criterion relatedness, whilst divergence refers to the extent to which measures of different traits tend to produce different results, regardless of

---

[18] Chapelle (1998) discusses three views on defining construct: trait perspective, behavioural perspective and interactional perspective. "Trait theorists [....] define construct in terms of the knowledge and fundamental process of the test taker [...] Behaviourists [...] define constructs with reference to the environmental conditions under which performance is observed [...] Interactionalists see performance as the result of traits, contextual features, and their interaction" (p. 34). From the trait perspective, these method effects (such as the effects of item or task on performance) are referred to as error resulting in inconsistent performance across the methods; from behaviourist or interactionalist perspectives, these method effects are not considered as error. Rather, the latter two perspectives expect context to influence performance.

whether the same test methods or different test methods (for example, direct writing tests vs. multiple-choice tests) are used. For the MTMM approach, therefore, more than one trait or construct and more than one method should be used.

Finally, testing consequences can also be used as evidence. According to Chapelle's (1998) review, Messick (1989) identifies two questions that should underlie the investigation of testing consequences: "what are the value implications of the interpretations made from testing? What are the social consequences of test use?" (p. 59). The studies (Meara, 1978, 1984) that investigate the testing consequences/impacts of two different test methods (for example, the word association test and the Y/N vocabulary recognition test for a vocabulary test) on the vocabulary acquisition theory and future vocabulary tests, and consequently on classroom teaching, can be examples for the study of testing consequence.

As discussed above, Messick (1989) suggested six types of evidence for validation. For the practical application of these validation methods, Fulcher's (2003) identification on the basis of these six types of evidence is helpful, as he divides the methods into two groups: quantitative methods and qualitative methods. Quantitative methods include correlation, factor analysis, the MTMM approach, Generalisability study (G-study) and multifaceted rasch analysis; whilst qualitative methods include expert judgement, questionnaires and interviews, discourse analysis and verbal protocol analysis. I will briefly explain some of them.

First, correlational evidence investigates the functional relationship between two measures, which are usually test scores: that is, whether one measure is associated with the other, and whether the two measures tend to vary in the same way (Bachman, 1990).[19] This correlational evidence is reported with a correlational coefficient from +1 to –1. In the case of an assessment of productive skills such as writing and speaking, this coefficient means either inter-rater reliability between two measures by two different raters (Fulcher, 2003) or intra-rater reliability between two measures by one rater on two different assessment occasions. These inter-rater and intra-rater reliability, or consistency of rating, are obviously the issue of reliability. However, consistent rating either between raters or within a rater requires consistent

[19] Agreement between raters is considered as reliability. All raters sometimes over-/under assess test-takers' performances, in which case their 'agreement' would be wrong. This shows that it could be problematic to assume that agreement between raters implies that the assessment is reliable.

interpretation of the rating scheme used to assess productive skills. Therefore, inter-rater and intra-rater reliability depend upon how the rater/raters interpret(s) the content of the scheme. Ingram (1990) mentions that the measure of inter-/intra-rater reliability can be said to be the measure of content-related evidence for construct validity of the rating scheme. This is why correlational evidence is included as one approach for construct validation.

Second, the G-study investigates "to what extent a test score, given under a specific set of conditions, will generalise across those conditions" (Fulcher, 2003: 211). In the case of a speaking test, for example, the G-study deals with the question of whether a score on the test can be generalised across certain facets such as raters and tasks. The first step is to specify the facets, such as raters and tasks, over which the score on the test should be generalisable. Next, the effects of each facet on the score are calculated using Analysis of Variance (ANOVA), with the specified facets taken as independent variables and the score on the test taken as a dependent variable (Fulcher, 2003).

Third, a questionnaire study can be very helpful in inquiring into the validity of a test or a rating scale. This method is meaningful in that it allows us to obtain the direct opinions and preferences of stakeholders in the testing situation (i.e., test-takers and teachers), something that is not possible with any quantitative method. Fulcher (1996c) used this method to enquire how test takers reacted to three types of task.[20]

Finally, think-aloud is a valuable method that can help reveal test-takers' or raters' introspection. With regard to think-aloud protocols, Green (1998) writes that:

> [t]he verbal protocols may show evidence of erroneous reasoning by raters, failure to note relevant features in a student's work that should be credited, or the use of criteria other than those recommended. Any one of these factors would reduce consistency of ratings. Importantly, this sort of information may not be directly inferred through the application of standard, quantitative approaches (p. 3).

Needless to say, this is based on the assumption that the rater who is doing think-aloud can verbalise his/her cognitive process (Fulcher, 2003) and that the rater's

---

[20] The responses to the questionnaire were analysed using statistical analysis such as factor analysis (Fulcher, 1996c).

verbalisations are an accurate and full record of what they were attending to as they performed the given task (Green, 1998).[21]

In this respect, it seems to me that diary study also can serve the same purpose as that of think-aloud, although it is not included as a validation method in Fulcher (2003). According to Bailey (1990) and Parkinson *et al.* (2003), diary study is usually used to investigate language teaching experiences, language learning experiences and student teachers' reactions to academic courses (e.g., Asher, 1983; Ho Fong Wan Kam, 1985). However, it could be used to investigate the information given by test-takers' or raters' introspection[22] on the process.

I have discussed validity and validation methods in detail. Even though a test or rating scale is justified as valid through the validation methods described above, there is one important point to be attended to: that validity is local rather than universal. As Davies and Elder (2005) state, citing Anastasi (1988), "the validity of a test cannot be reported in general terms. No test can be said to have 'high' or 'low' validity in the abstract. Its validity must be established with reference to the particular use for which the test is being considered" (p. 139). In other words, although a test or a rating scale is valid for a specific use and context, it cannot be said to be universally or generally valid.

This section has presented a detailed discussion of validity, which is generally the main concern in all fields of language testing. Given that validity has come to be seen as increasingly important in the assessment of writing as well, I will narrow the discussion down to the validity of writing assessment. With regard to the issue of validity in writing assessments, the variables that might affect their validity have been identified and investigated in previous studies: writing task, scoring procedure, test-taker (Hamp-Lyons 1990; Huot 1990a, 1990b; Weigle 2002) and the uses or consequences of tests (Shohamy, 1997). These variables will be discussed in detail in the next section.

### 3.3.2 Factors affecting the validity of writing tests

[21] There is a criticism on this assumption (e.g., Glendinning & Howard, 2001).

[22] The information obtained by diary study is usually retrospective, whilst information from the latter is concurrent.

There are many studies that deal with the factors affecting the validity of a writing test (e.g., Brossell, 1986; Carlson & Bridgeman, 1986; Hamp-Lyons, 1990, 2003; Hamp-Lyons & Kroll, 1996, 1997; Huot, 1990a, 1990b; Shohamy, 1997; Storch, 1993; Weigle 2002). They usually identify four variables: the writing task variable, the scoring procedure variable, the test-taker variable and the test consequence variable.

First, regarding the writing task variable, many researchers (e.g., Choi, 2000; Pollitt & Hutchinson, 1987) have investigated whether a writing task variable affects the test-takers' performance in writing tests, and have generally found that it does. For example, Choi (2000) asked Korean subjects to do two kinds of writing tasks: an information transfer task and a free writing task. The former required the test-taker to give their own interpretation of non-verbal input. This had a more subject-related vocabulary. The latter asked them to write about the advantages and disadvantages of the computer. The results showed that test-takers wrote more fluently in the free writing task than in the information transfer task.

It has also been found that the validity of a task might be affected by various other factors: purpose, audience, mode of discourse, length of time, topic, rhetorical specification and the use of pen and pencil or word processor, etc. Cohen (1994) pointed out that aside from the issue of rating scales, the main issue with writing assessment is about the prompts for writing, which include purpose, audience, mode of discourse, length of time, topic and rhetorical specification and so on. He contends that as test-takers' scripts are assessed on the adequacy of their written responses to one or more prompts, these prompts need to be written in the most writer-friendly way. Kroll and Reid (1994) also note that to achieve its purpose of ranking, sorting, or placing students, a well developed prompt should reflect various variables. Otherwise, it might result in unfinished, unfocused or rambling essays which show that test-takers have misunderstood the question.

Of the elements in the prompts, topic is the most controversial in terms of its potential influence over a test-taker's performance. A number of studies (e.g., Brossell, 1986; Brown *et al.*, 1991; Reid, 1990; Tedick, 1990) have investigated the effect of topic on the writing performance of college students. Some of them argue that test-takers' performance varies according to the degree of familiarity with a

given topic on which they are asked to write, whilst others contend that they do not show any differences on performance across various types of topic.

Brown *et al.* (1991), for example, investigated the degree to which topic types and individual prompts affect the performance of college freshmen taking the Manoa Writing Placement Examination. Their study covered 3,452 freshmen subjects, who were each asked to write on two prompts. One prompt was to write an analytic essay after reading approximately one and a half pages of prose, the other was to write an essay based on personal experience. An identical scoring method (holistic scoring) was employed to rate the essays, and the raters were trained prior to the scoring session. The results showed a significant correlation between the prompt sets and performance. This correlation is also observed in Tedick's study (1990) which shows that the ESL students in the study performed better on a field-specific topic than a general topic and that the former discriminated between different levels of writing proficiency better than the latter.

Meanwhile, Read (1990) suggests the extent of information provided within a prompt as a factor which might affect validity of a task. He classifies writing tasks into three types according to the amount of preparation or guidance that the test-takers are given: independent tasks, guided tasks and experience tasks. In the case of independent tasks, test-takers are expected to write on a topic without any guidance. In this kind of task, it is assumed that all students have the background knowledge necessary to do the task. With guided tasks, students are provided with guidance on the content of what they are writing, in the form of a table, a picture or linguistic material. For experience tasks, students are given the opportunity to acquire relevant knowledge on a given topic prior to the writing session. With this classification of tasks, Read finds that the second and third types are preferred to the first one in EAP writing tests (e.g., ELTS, TEEP and TWE). It is believed that these two types of task are more valid for writing assessment than the first one, which is likely to assess test-takers' background knowledge as well as writing ability.

Although there are many studies on tasks for writing assessment, and some correlations were found between the task variable and performance on the task, there are, as Ruth (1982) confesses, "few answers to the kinds of questions that researchers are asking about the nature of the effects of writing prompts upon performance" (p.

63, cited in Brossell, 1986).

Secondly, with regard to the scoring procedure variable, this is divided into two subcategories: a scoring scheme variable and a rater variable. Assessment categories, which are either holistically or analytically included in a scoring scheme, directly indicate which qualities are intended to be measured through the writing test. Consequently, depending on which assessment categories are included in the scheme, this might in turn affect the degree to which the test can be said to be valid. Furthermore, as Cohen (1994) points out, since individual scales may include categories on which it is difficult to make qualitative judgements – such as the level of coherence in an essay or the acceptability of its style – applying scales that include such categories and achieving validity is likely to be difficult.

Although these qualities are reflected in a scoring scheme and all raters in a writing assessment may be expected to use it identically, whether the qualities are actually measured as intended is another question, since the actual use of the scheme depends on the raters, who will vary in terms of their understanding and application of the scheme (even if they are trained in the scoring procedure). Weigle (2002) explains the reasons for this variation as follows:

> Raters bring their own background, experiences and values to the assessment of writing and while training can help bring raters to a temporary agreement on a set of common standards, research has consistently shown that raters will never be in complete agreement on writing score (p. 72).

Therefore, it is likely that the validity of the writing assessment will be affected by the raters' understanding and application of the scheme, as well as their basic understanding of the purpose and importance of the assessment (in actual rating, validity of rating might be also affected by contextual variables such as rating timing and behaviour of trainers). In light of this, it can be said that a lack of reliability in writing assessment scoring may be caused by a lack of consistency itself, but simultaneously and more seriously by a lack of validity. Hamp-Lyons (1990) explains this as follows:

> ...nor can they [raters] agree on the specific qualities in essays that make them good, worse, or worst..... This is a validity problem as well as a reliability problem.

> Since we as researchers cannot consistently agree with each other when assessing the same writing samples or even sometimes with our own judgements about the same samples made on different occasions, we cannot be looking at the same thing. That is, we do not share a construct of writing quality (p. 80).

Third, the test-taker variable has attracted little attention from language testing theorists and practitioners (Hamp-Lyons & Kroll, 1996). This is, however, a critical variable which might affect the validity of writing assessments in a similar way to the variables mentioned above. This variable is mainly to do with writers' topic interpretation. Writers might vary in every step of attending to and understanding a given identical prompt. Alderson and Clapham (1995) note that test-takers often interpret test items differently from the way intended by the test developers. As a result, some of them might do the task in a different manner from that which is intended, because "giving and responding to an assignment is an act of negotiation" (Ackerman, 1990: 96). Hamp-Lyons and Kroll (1997: 21) agree to this, citing Labov (1969) who said that it is "absurd to believe that an identical 'stimulus' is obtained by asking everyone the same 'question'" (p. 108).

In addition to this aspect of topic interpretation, writer apprehension is another important part of the test-taker variable. For this, Brossell (1986) reviews Faigley *et al.* (1981) who investigated whether highly apprehensive writers would perform differently from less apprehensive writers in both standardised tests of writing-related skills and in two essay tests. They found that there were significant effects for apprehension with personal narrative essays, but no effects with argumentative essays. The researchers suggested that this was because apprehensive writers might be more anxious about expressing personal feelings and experiences than about arguing objectively.

Finally, I consider the test consequence variable, which has been given less attention than factors associated with the test itself mentioned above. Shohamy (1997) argues that the test consequence variable is one of the major sources of variability that might affect the validity of tests. She writes that "research on the uses of language tests reveals that [some] tests are used for different purposes from those they were intended for and thus can be viewed as unethical and unfair" (p.343). In one example, she cites a test of English as a foreign language used in Israel. She mentions that after conducting this new type of test, the national education database

collected records of every student on the test, and that language inspectors then used the tests as a vehicle to trigger changes in the education system (without changing the curriculum and systems for teacher training) and to achieve bureaucratic agendas, deviating from the intended target of simply measuring students' proficiency. She insists that the use of tests like this is unethical and unfair: that it does affect the validity of the test as a consequence.

Obviously, writing tests need to be valid in terms of this test consequence variable. However, as Davies (1997) maintains, it is not possible to take account of all the possible social consequences of tests. He claims that we can be responsible for limited and predictable social consequences, and aim to prevent internal bias in tests and be willing to take account of test fairness.

### 3.4 Alternative writing assessment: portfolio assessment

As we have seen, there are various factors which might affect test-takers' performance and in turn the validity of testing. To achieve control over all variables, however, is an unlikely prospect, especially in timed impromptu tests. Furthermore, "a single sample collected under timed impromptu conditions does not reflect [...] concerns with processes, strategies, growth..." (Hamp-Lyons, 1991a: 14). The timed impromptu test restricts the multi-fold domain of writing to narrow categories, and because of the time limit and test context, this kind of test strips natural context from writing. As a result, timed impromptu tests yield limited information on the writing ability of test-takers (White, 1995).

These limitations of timed impromptu tests even in terms of validity, which is nonetheless the main virtue of direct tests over indirect tests, have been pointed out, in my view, due to the shift from a focus on raters to a focus on test-takers. When I summarise the history of writing assessment, the main concern in the past was how raters could assess "well" – i.e. reliably. This led to the introduction of indirect writing tests, and then when direct tests came back into favour again, attention was paid to the reliability of holistic/analytic scoring in direct writing tests. After the 'excessive focus on reliability' passed by, testing researchers turned their attention to validity, and thereby to test-takers. Efforts to assess well on behalf of "*test-takers*" now started to be made. Another development was that the revising process was

recognised as a necessary step in the writing process, and it was proposed that revising ability should be included in writing assessment. Similarly, in order to help test-takers demonstrate their ability in writing tests as well and as fully as possible, alternative assessments, especially portfolio assessment attracted the attention of testing researchers.

Portfolio assessment is defined as a collection of reflective comments by a writer and "texts written for different purposes over a period of time" (Weigle, 2002: 198). Murphy and Smith (1991) refer to portfolio as a 'process-folio', in that it includes multiple drafts. Reflection and selection are considered to be essential processes in compiling a portfolio, because the entries are selected by the writer. Although the types of portfolio can vary, the best portfolio should consist of writing "from different points over the course or year and take into account both growth and excellence" (Hamp-Lyons, 1991b: 262).

This portfolio assessment is said to have many advantages over timed impromptu writing tests. First of all, the most obvious advantage of the portfolio assessment is the validity of the inferences made by raters about the writer's ability. As this assessment of writing ability is not based on a single piece of work, but on multiple pieces of work on various kinds of tasks/topics, for different audiences and in various contexts, raters can be confident in generalising the level of writing ability from the assessment results of a portfolio. Second, as portfolios strongly support pedagogies that include multidrafting, revision, peer review, collaborative learning and reflective writing, portfolio assessment appeals to teachers of L2 writing. In addition, as portfolio assessment has a direct connection between what is taught and what is assessed, evaluation can be matched with the teaching objectives. Third, encouraging students to build multi-genre portfolios can help them understand how texts are organised differently to express particular purposes depending on the genre. Fourth, portfolio assessment has authenticity in that this kind of test can be designed to include writing samples which were written for a real purpose, not for the purpose of evaluation. Finally, portfolio assessment has a positive backwash effect on teaching and learning, in that it can promote students' autonomy in learning by giving them opportunities to develop their self-awareness through the process of reflection on their own writing, and providing opportunities for students to have

ownership of their writing through the process of selecting the entries (Hyland, 2003; Miller, 1995).

Judged on its advantages alone, portfolio assessment would be vigorously pursued and conducted within the discipline of writing assessment. However, there are obstacles to be overcome before it can be employed as a means of assessment at national as well as classroom level, (Hyland, 2003; Miller, 1995). These mainly concern the scoring procedure. Scoring a portfolio is harder than assessing a single piece of writing. What is worse, a portfolio is a compilation of heterogeneous examples of writing, and it is extremely difficult to select representative pieces that best demonstrate a student's writing ability in a portfolio, and many decisions have to be made before the selection. Accordingly, the following questions regarding portfolio assessments need to be resolved: how can raters reliably assess through samples in a portfolio? To what extent can raters assess writing samples without being affected by their impression of the first piece in the portfolio (i.e., without the halo effect[23])? How many human resources (raters) and how much time is available for portfolio assessment? How can portfolio assessments be used in a highly constrained curriculum where limited time and resources are available? Hyland (2003) summarises the advantages and disadvantages of portfolio assessment mentioned above in a table shown below as Table 3.3.

Table 3.3 The advantages and disadvantages of portfolio assessments (Hyland, 2003: 236)

| Advantages | Disadvantages |
|---|---|
| Represents programme goals | Produces heavy workload for teachers |
| Reflects progress over time, genres, and conditions | May encourage "teaching the portfolio" |
| More broad, comprehensive, and fair than exams | Difficult to compare tasks set by different teachers |
| Closely related to teaching and students' abilities | Difficult to assign a single grade to varied collection |
| Students see portfolio as a record of progress | Problems with plagiarism or outside assistance |
| Focuses on multi-drafting, feedback, revision, etc. | Problems with reliability across raters |
| Assignments build on each other and show genre sets | |
| Allows different selection and assessment criteria | |
| Students reflect on their improvement and weaknesses | |

[Originally Table 8.5]

In sum, as Hyland (2003) points out, "portfolios do not necessarily bring greater

[23] Although it is generally agreed that the halo effect needs to be avoided, in fact it is to some extent inevitable (Storch, 1993).

accuracy to assessment, but they do promote a greater awareness of what good writing might be and how it might be best achieved" (p.239). White (1994) also believes that portfolios have both considerable advantages, such as validity and a close relationship between assessment and teaching and problems such as the reading time required for teachers, who are already overburdened with many classes and students.

In this chapter, I have discussed different types of writing tests (direct tests, indirect tests and portfolio assessments). They are not always good. The preferred test type therefore still depends on the assessment situation and context, as with the choice between direct and indirect tests. Timed impromptu tests do not assess the test-takers' revising ability, but in terms of time, cost and resources for scoring, it can be meaningful to assess their ability to write a first draft. These impromptu tests can also be appropriate in the case of determining whether students are ready for a freshman composition course in a college or university. On the other hand, they may be inappropriate in assessing advanced or graduating students, who should be able to use sources intelligently to support their ideas and demonstrate their understanding of a topic (White, 1995).

Therefore, it might be suitable to conclude this review of different test types with White's (1995) comment that:

> No assessment device is good or bad in itself but only in context. Only when we know what we are seeking to discover, can we claim that a particular kind of assessment is appropriate or not (p. 34).

## 3.5 Conclusions

In conclusion, none of the test types – direct tests, indirect tests, portfolio assessments – are good in absolute terms. This is partly because the construct, i.e. writing ability, is not simple to measure, as mentioned in Chapter Two, and partly because the choice of tests should depend on the purpose, context and resources available to conduct writing tests.

For the direction of future writing assessments, Hamp-Lyons and Kroll (1996) suggest that in order to find out what the writing abilities of test-takers are really like, a test should be designed that at some level acknowledges the writing processes on

which all writing products depend. However, this will not be easy to achieve, for as White (1995) summarises, "all measurement of complex [writing] ability [including the aspect of writing process] is approximate, open to question and difficult to accomplish" (p. 43). Hamp-Lyons and Kroll (1996) also share this opinion, arguing that "given that 'writing is a multidimensional, situational construct that fluctuates in a wide variety of contexts' (Greenberg, 1992: 18), it is perhaps inevitable that creating the best of all possible tests and testing situations remains a constantly shifting balancing act" (p. 68).

In this chapter I have discussed writing assessment in general, as a background to the rating scale which I have developed. Of the types of writing assessment discussed in this chapter, the type for which the rating scale was designed was direct assessment. Given that the present concern is validity, and that certain variables may affect the validity of direct writing tests, these need to be taken into account when the tests are considered. In the next chapter I will discuss the scoring criteria used to make the assessments.

# CHAPTER FOUR. SCORING CRITERIA

## 4.1 Introduction

For rating schemes there are two methods available. One uses a formal rating scale such as a holistic scale, an analytic scale, a primary-trait scale or a multiple-trait scale; and the other is subjective holistic scoring, which does not use any formal rating scales. Given the recent tendency to use a rating scale (Alderson, 1991) and the purpose of this study, I will focus on the former in this chapter, and discuss aspects which need to be considered before a rating scale development.

In section 4.2 I will discuss the history of rating scales, in section 4.3 the nature of rating scales, in section 4.4 the classification of rating scales, and in section 4.5 two approaches to rating scale development. In section 4.6 I will deal with the other sub-variable, the rater variable, focusing on the effect of raters' backgrounds and training on their judgement.

## 4.2 The history of rating scales

Systematic concern with the development and use of rating scales in second language teaching dates back to the 1970s, as rated tasks have increasingly been used in place of objective items in second language tests (Upshur & Turner, 1995). Alderson (1991) reviews that since the development of the Foreign Service Institute (FSI) scale in the US (for oral proficiency assessment),[24] various rating scales have been developed following the characteristics of the FSI, one of which is to refer to the performance of native speakers for the top level of the scale: the ILR (Interagency Language Roundtable) scale (for oral tests only), the ACTFL guidelines and the Australian Second Language Proficiency Ratings (ASLPR) scales.[25] In addition to these, other rating scales have been developed, including the ESL Composition Profile by Jacobs *et al.* (1981), the TEEP scale by Weir (1990), the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1991b) and the TWE scale by ETS. Following the tradition of the FSI, these scales try to describe the changes as

---

[24] Clark and Clifford (1988) introduce both the background and procedure of development of the FSI scale in detail.

[25] The name ASLPR was changed to ISLPR (International Second Language Proficiency Ratings) in 1997 (http://www.gu.edu.au/centre/call).

interlanguage develops on the continuum from zero to native-like, so that they provide a comprehensive picture of language behaviour (Ingram, 1995).

In the UK, the ELTS (English Language Testing Service) scales were developed in 1980. Afterwards, as the IELTS, the CPE, the CAE, and the FCE were developed for testing proficiency in English as a second language, rating scales specific to these tests were developed accordingly.

### 4.3 The nature of rating scales

To understand the nature of a rating scale, I will approach it from various viewpoints. To start with, there is its physical appearance. A rating scale usually has two axes: one is a horizontal axis for assessment categories, and the other is a vertical axis for bands or levels. Thus, the rating scale concerns itself with measurement of the descriptions.

This feature is revealed through the definitions of rating scales. According to North's (1995, 2000b) review of the definitions, it is "a hierarchical sequence of performance ranges" (Galloway, 1987: 27), or "characteristic profiles of the kinds and levels of performance which can be expected of representative learners at different stages" (Trim, 1978: 6). Ingram (1995), in a similar vein, defines rating scales as "a graduated series of descriptions of language behaviour or of selected aspects of it" (p. 17).

All these definitions reflect superficial or physical features of the rating scale. To better comprehend the nature of rating scales, it would be helpful to consider descriptions and definitions based on construct as well. Storch (1993) defines the rating scale as "a *construct* [...] encoded in the wording of the rating scale" and "an implicit view of language proficiency" (p.22) [emphasis added]. Hamilton *et al.* (1993) agree:

> The terms in which the rating scales descriptors are couched are important, because they constitute implicit definitions of the construct on which the test is based (p. 2).

In these definitions, a rating scale can be understood as a construct of assessment.

To understand the nature of a rating scale, consideration of the relationship between the scale and the test to which it is related is also desirable. In a word, the

relationship between the two is such that even though assessment is carried out using the rating scale, the latter is not a measure on its own. That is, "a scale is not an instrument but a sort of metaphor to inform a judgement" (Davies, 1995: 9). Ingram (1995) and Hill (1995) agree with this, and Ingram (1995) also takes the view that "scales are used to explicate and assist in the interpretation of test results while tests of a variety of different types may be designed specifically to assign learners to levels on a scale" (p. 15).

With regard to the nature of the rating scale in terms of the principles of rating scale development, it is context-specific. As Hamp-Lyons (1991) and Storch (1993) emphasise, just as a test should be developed to reflect the context and purpose in which and for which it is used, a rating scheme needs to have input from the programme in which it is used. Therefore, in order for it properly to function, some questions need to be asked beforehand, such as what features should be considered important in the given context; what it should be like in order to suit the specific group of learners; and what is the purpose of assessment.

To summarise, a rating scale is not a test but a tool to assist and interpret the rater's judgement. It should refer to the constructs, describe differences between proficiency levels and ideally be context-specific. The next section continues the investigation of rating scales by looking at their types and classification.

### 4.4 Classification of rating scales

As we have seen in section 4.2, there are many rating scales. They can be classified according to various criteria, such as the approach through which they have been developed and the purpose for which they have been developed. According to Bachman (1990), there are two approaches to rating scale construction: the real-life or behavioural approach, and the interactive-ability approach. In the former, the rating scale is intended to represent a picture of what a test-taker at a particular level can do in the real world. Many scales of proficiency are based on this approach, particularly rating scales for vocation-related performance. This kind of scale is based on a notion of increasing complexity in the tasks. According to North (2000b), the ten-band Eurocentres global scale, of which an extract is given below, exemplifies this approach.

Band 7
Can express ideas and opinions clearly on a wide range of topics, and understand and exchange information reliably. Has an active command of the essentials of the language. Can communicate competently and independently in many professional as well as personal contexts.

Band 6
Can understand information on topics of interest in unsimplified but straightforward language and can find different ways of formulating what he or she wants to express. Has assimilated the essentials of the language. Can communicate competently in many professional as well as personal contexts.

Band 5
Can understand extensive simple information encountered in everyday situations and maintain conversation and discussion on topics of interest. Can exploit a wide range of simple language flexibly to express much of what he or she wants to. Can communicate adequately in routine professional contexts.

(from Eurocentres global scale, quoted from North, 2000b: 9-10)

However, even if a scale is developed by the real-life approach, as Ingram (1995) points out, it cannot encompass all real-life language behaviour, but "is necessarily selective and suggestive of real life language behaviour" (p. 13). He explains this point as follows:

> The complexity of language and its variations from situation to situation according to who is using it, to whom, in what medium, in what location, for what purposes, and about what topics, all means that, if the scale descriptors actually attempted to match real life language performance, they would be unmanageable (p. 13).

On the other hand, in the interactive-ability approach the rating scale is constructed to describe constructs: aspects of the test-taker's performance in a particular test. North (2000b) suggests the rating scale by Upshur and Turner (1995) as an example of this category (see section 4.5.3 for their rating scale).

Alderson (1991) suggests that rating scales can be classified according to the purposes for which they are developed and used. First, they can "serve to describe levels of performance" (p. 72). The ILR scale below is an example of this type, which is called a user-oriented scale.

> able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations, including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations

> on nontechnical subjects (i.e. topics which require no specialised knowledge) and has a speaking vocabulary sufficient to respond simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar
>
> (S-2 in Speaking part of the ILR, quoted from Alderson, 1991: 72)

Second, they can serve to "provide guidance for assessors who are rating performances" (p. 73). Therefore these kinds of rating scales are called assessor-oriented scales. The ASLPR (ISLPR) can be taken as an example of this.

> Can write a personal letter on simple everyday topics or a simple report on an everyday event. Can write to order goods, book a room, or to carry out other uncomplicated and routine tasks. Can fill out most forms regularly encountered in everyday life (e.g. health insurance, unemployment registration, passport application, etc.)
>
> (W;2 in writing part of the ASLPR, quoted from Alderson 1991: 73)

Finally, they can serve to "provide guidelines for test constructors--a set of specifications, if you will, of the sorts of texts, tasks and items that a test appropriate for a given level of student should contain" (p. 73). The rating scales with this function are called constructor-oriented scales. The ACTFL Guidelines are an example of this type.

> Sufficient comprehension to read simple authentic printed material or edited textual material within a familiar context. Can read uncomplicated but authentic prose on familiar subjects containing description and narration such as news items describing frequently occurring events, simple biographic information, social notices, and standard business letters. Can read edited texts such as prose fiction and contemporary culture. The prose is predominantly written in familiar sentence patterns. Can follow essential points of written discussion at level of main ideas and some supporting ones with topics in a field of interest or where background exists. Some misunderstanding. Able to read the facts but cannot draw inferences.
>
> (Advanced level in Reading part in the ACTFL guidelines, quoted from Alderson, 1991: 73-4)

In addition to these three categories of rating scales (user-oriented, assessor-oriented and constructor-oriented[26]), Pollitt and Murray (1996) add another category: a diagnosis-oriented scale. They found that raters paid attention to different aspects depending on the different levels of test-takers during assessment. However, many

[26] More examples of each type of rating scale can be seen in North (2000a).

professed assessor-oriented rating scales are constructed with a full grid, that is, they allow all possible combinations of levels (vertical axes) and categories (horizontal axes), and so do not reflect this aspect of assessors' rating behaviour. Researchers argue that if a rating scale is to be truely assessor-oriented, it should not have such a detailed description of a particular feature if the feature is not salient for a level. They argue, therefore, that this kind of assessor-oriented rating scale is not really for assessors since its use may complicate the rater's task rather than aid it. Instead, according to them, this type of scale is for the purpose of diagnosis. Therefore, they refer to it as a diagnosis-oriented rating scale instead of an assessor-oriented rating scale, while referring to a rating scale which reflects a rater's behaviour as assessor-oriented.

Ingram and Wylie (1991) and Ingram (1995) also classify rating scales. They classify nine pairs of contrasts in rating scales according to three criteria: what they intend to measure, how they are constructed, and what criteria they contain. The contrasts and their characteristics include the following:

1. Whole vs. Part

   depending on whether a scale has to do with either the whole span of proficiency development or only a part of it

2. Serial vs. Threshold

   depending on whether a series of intermediate points between two bands are provided in a scale or a scale is described in a threshold level with a more cursory description

3. General vs. Specific purpose

   depending on whether a scale describes general language proficiency or language proficiency in some specific area

4. Task-only vs. Total or underlying behaviour

   depending on whether a scale graduates some specified task-related proficiency levels or total or underlying behaviour regardless of tasks

5. Proficiency vs. Course achievement

   depending on whether a scale describes general proficiency or specific course-related performance

6. Macroskill-specific vs. Overall

depending on whether a scale describes language behaviour in one or more of the macroskills separately or combined, or describes general language behaviour in a way that is supposed to relate to all of the macroskills

7. Absolute vs. Global

depending on whether all criteria within each band in a scale must be absolutely fulfilled before a learner is rated at that band, or whether it seeks to provide a global picture of language behaviour

8. Analytic vs. Holistic

depending on whether a scale provides a detailed and specific statement or more general statement about general development in language learning

9. Empirical vs. Washback effect

depending on whether a scale describes language behaviour as observed or what is considered to be desirable for success in a course

While various suggestions have been made for the classification of rating scales along such lines, one of the most common classifications in the nine pairs may be holistic scales and analytic scales (multiple-trait scales for Hamp-Lyons (1991b)). Therefore, I will discuss these scales further.

For holistic scales, the ACTFL guidelines, the ASLPR (ISLPR) and the TWE scoring guide can be taken as examples. In these scales, assessment categories such as content, organisation, vocabulary, grammar and mechanics are not differentiated, but are merged into a few or several descriptors in a band. A rater who is using the holistic scale is invited to read a script quickly and to judge it, based on his/her overall impression, using the holistic scale. In holistic scoring, the result of scoring is represented in a single score.

For analytic scales, the ESL Composition Profile, the TEEP scale and the Michigan Writing Assessment Scoring Guide can be taken as examples.[27] Of these, the ESL Composition Profile by Jacobs *et al.* (1981) was the first analytic scale for L2 writing assessment. In these scales, assessment categories are independently classified, and as a result it is possible to have a script rated on each category. Multiple scores on the categories are either reported separately[28]or added up.[29]

---

[27] Hamp-Lyons (1991b) refers to this as multiple-trait scale.

[28] If a construct in a test or rating scale is assumed to be unidimensional, rating results can be reported

Analytic scoring using these analytic scales can thus provide detailed and diagnostic information about a test-taker's level of writing, compared with holistic scoring where he/she is given a single score for his/her writing. Additionally, analytic scales have been shown to be more reliable than holistic scales (Hamp-Lyons, 1991b; Hill & Storch, 1994; Storch, 1993).

In addition to these two scales, the primary-trait scale is found in Lloyd-Jones (1977, cited in Weigle, 2002). This kind of scale is developed with respect to a specific type of writing. A primary trait that is identified as important for successful writing in each test is defined by the test constructors, and the levels of success in the trait is determined. The trait will vary depending on the topics and writing type (Cohen, 1994). The construct of writing ability is, accordingly, defined very narrowly compared with either holistic scoring or analytic scoring, which are not limited to a specific topic and writing type. Since a rating scale must be developed for each writing task, this is time and labour intensive. Furthermore, as Hamp-Lyons (1991b) claims, since verbal reports by raters revealed that they may have difficulty focusing exclusively on the one specific trait, and may inadvertently consider other traits during rating, this may threaten the reliability and validity of assessment.

When it comes to the question of which is preferred of these three types of scale, discussion is usually restricted to holistic scales/scoring and analytic scales/scoring, just as Ingram and his colleague did above. There appear to be three reasons for this. First, the other kind of scale, primary trait, is less practical and popular than the holistic scale and the analytic scale because of the drawbacks discussed above. Second, both holistic scoring and analytic scoring have apparently contributed more to the popularity of direct tests in L2 writing assessments. The development of holistic scoring has been one of the biggest breakthroughs in direct tests, and analytic scoring has been adopted as a means of remedying any deficiencies in holistic scoring. Third, since the holistic scale and the analytic scale are the complete opposite of each other in terms of both form and rating procedure, researchers' discussion has tended to focus on these two contrasting categories. Therefore, I will

---

in a single score. However, if the construct is assumed to be multidimensional, it is better to report scores in profile (Fulcher, 1999).

[29] Since it is not practical to report in multiple numbers for a script, it is desirable to add up the multiple scores on the assessment categories (Hamp-Lyons & Prochnow, 1991).

discuss these two types of scales/scoring in more detail.

Many studies have been carried out, either to advocate or criticise holistic scoring. One of its advocates, Huot (1990a), claims that even though holistic scoring is sometimes considered invalid, this is not the case. According to him, it is the emphasis on the reliability of holistic scoring that has made it vulnerable to criticism, especially in the current situation where validity is considered the most important quality in measurement. Furthermore, reviewing previous studies to determine which features dominate holistic scoring, he concluded that holistic raters are most influenced by the content and organisation of a test-taker's writing, dispelling the assumption that holistic scoring correlates with appearance and length of writing and thus lacks validity.

Charney (1984), another advocate of holistic scoring, also examines its validity. Charney asserts that:

> It might be argued that the evaluation of writing ability should take the individual's writing *process* into account, and that product-based methods of assessment, such as holistic ratings, are therefore invalid *a priori*. This kind of argument assumes that if holistic ratings are invalid for one purpose, namely for assessing the student's writing process, then they are invalid for all purposes. [..] Product-based evaluations, including both quantitative and qualitative methods, do not produce diagnoses [...]. Instead they produce summative statistics which compare the abilities of individuals or of groups of writers (p. 68). [originally emphasised]

Matthews (1990) is also generally opposed to analytic scoring. She mentions that in speaking assessments as well as writing assessments, analytic scoring may threaten to complicate administration and jeopardise validity.

As we can see, the advocates of holistic scoring emphasise that it is less time-consuming than analytic scoring, equally reliable (Cumming, 1990) and gives valid information on writing ability. Nevertheless, holistic scoring has been accused of having several shortcomings. The first is that it fails to reveal how a rater arrives at a final/global rating using the holistic scale. As can be seen in many studies (e.g. O'Loughlin, 1993; Shi, 2001), raters differ in which assessment features they give most weight to, and categories such as content and organisation influence each rater differently. However, this is masked in holistic scoring. Consequently, holistic scores cannot be explained adequately to anyone but the rater him/herself (Hamp-Lyons,

1991b).

Second, to reduce complex writing ability to a single number is too "reductive" (Hamp-Lyons, 1991b: 244), meaning that it cannot indicate subtle differences in the learner's writing.

Finally, holistic raters are asked to attend to all writing features, such as content, organisation, style, grammar, mechanics and vocabulary. It is recognised even amongst the advocates of holistic scoring that rating so many writing qualities may be overwhelming for raters.[30]

As a result of this criticism of holistic scoring, various analytic scales have been proposed (Hamp-Lyons, 1990, 1991a, 1991b; Sasaki & Hirose, 1999; Shi, 2001). The advocates of analytic scoring argue first, that it can produce more reliable results than holistic scoring, and second, that it can give diagnostic and detailed information on writing ability. As Hamp-Lyons (1996) and Hamp-Lyons and Prochnow (1991) note, writing is complex, i.e. multidimensional. Thus, it is not reasonable to try to capture the strengths and weaknesses of a test-taker's writing in a single number, and this is especially true of L2 writing. Advocates argue, therefore, that when ratings are given in the form of a profile through analytic scoring rather than in the form of a single score, both L2 learners and teachers may benefit from this information and use it to improve L2 writing ability. Given the fact that writing in L2 does not develop at the same speed in every category, this is considered the greatest advantage in analytic scoring, because the assessment results show learners' strengths/weaknesses (Clark, 1985). Shi (2001) also addresses this point, as follows:

> ....holistic scoring was not effective in distinguishing subtle differences in students' writing performances, nor was it an effective way to detect differences between NES and NNS teacher raters. Analytic ratings or qualitative evaluation, as the present study implies, might be preferable to holistic scoring to assess accurately the quality of L2 writing... (p. 316-7).

In addition, as Cumming (1990) mentions, since an analytic scale makes specific

---

[30] To overcome this drawback, Greenhalgh and Townsend (1981) suggest an alternative holistic scoring method: focused holistic scoring. It is called 'holistic' "because it considers the total piece of writing; it is 'focused' because it evaluates writing in terms of pre-defined criteria" (p. 812). According to them, the use of this focused holistic rating scale helps both teachers and learners to pay more attention to the content of writing rather than the mere surface features of writing.

aspects of learners' scripts explicit, it could guide the decision-making process of novice raters during assessment. Furthermore, it could help even experienced raters whose decision-making varies from person to person in holistic scoring.

However, there are also disadvantages with analytic scoring. For one thing, it takes more time to rate a test-taker's work with analytic scoring than with holistic scoring, because raters need to read a script more than a couple of times in order to judge it in multiple assessment categories. This is its greatest defect (especially in a large-scale assessment). Next, the halo effect may be found in analytic scoring. That is, judgement in one assessment category, such as grammar, might affect judgement in others, such as content and organisation.[31] This means that the analytic scale is not being used as it was intended to be used – to provide more detailed information on each category; also, that judgements are blurred (Hill & Storch, 1994). Additionally, although it is assumed that analytic scoring provides more reliable information on writing ability than holistic scoring, this may not be the case. Hill and Storch (1994) had scripts double marked and had raters trained to achieve acceptable levels of inter-rater reliability. Nonetheless, analytic scoring failed to achieve equally acceptable levels of reliability across the assessment categories. High inter-rater reliabilities could be achieved for Grammar and Vocabulary, but levels were lower for Communicative Quality and Argument. The authors interpreted these results as partly due to the fact that the two former categories are more concrete and easier to apply than the latter two. Hwang (1930) also found that the analytic rating scale used in his study (the Hudelson English Composition Scale) did not prove to be as reliable as expected. Given these results, it can be concluded that analytic scoring also has defects.

## 4.5 Approaches to rating scale development

### 4.5.1 Overview

Turner and Upshur (2002) divide the approaches to rating scale construction into three types; the *a priori* approach, the data-based approach and the learning objectives-based approach. The first involves identifying the features of performance

[31] According to the literature on the halo effect, grammatical accuracy or vocabulary usage is often likely to be related to the overall score (Hill & Storch, 1994).

at each level on the basis of a theory of the development of L2 ability. In the second, a scale is empirically developed on the basis of, for example, the developers' description of differences in a sample performance of L2 learners or direct analysis of a sample performance of L2 learners. The last type of scale development is done on the basis of analysis of the sequence of learning objectives in a specific L2 course. Of these, the first two approaches are discussed most, apparently because they are diametrically opposed to each other in that they are *a priori* vs. empirical, or theory vs. data. Thus, I will also focus on these two approaches for the discussion of rating scale development.

### 4.5.2 The *a priori* approach

The *a priori* approach to rating scale development entails developing a rating scale on the basis of language learning theory without using any empirical data. It is also called the theory-based approach. According to Ingram (1995), there are two ways to develop a rating scale according to this approach: one is proficiency-based, the other is task-based. The former seeks to describe how interlanguage develops from zero to native-like, that is, it assumes a span of the development of interlanguage. This is, therefore, concerned not only with identifying what a learner can do but also with sequencing the identification.

On the other hand, the task-based approach seeks to identify the tasks or targets which constitute the activities and purposes of a given context, for example, a specific occupation, specific curriculum or course. Put another way, the activities a learner must be capable of in order to succeed in a given context are identified, and the rating scale is developed on the basis of this. Rating scale developers using this approach, therefore, need to observe the activities and contexts in question before specifying the tasks.

These theory-based/*a priori* approaches to rating scale construction and the resulting scales have been criticised on several counts. To begin with, they are based on *assumptions* rather than observed reality. Clark (1985) asserts that the definitions of the rating scales mentioned in section 4.3 reveal them to be "descriptions of expected outcomes, or impressionistic etchings of what proficiency *might* look like as one moves through hypothetical points or levels on a developmental continuum"

(p. 348). He points out that although the descriptors in the rating scales look logical, they never reflect actual linguistic behaviour because the theory-based rating scales have not been developed empirically (North, 1994). According to Scarino (1995), "they represent teachers' conceptions of what students are expected to be able to do as opposed to what they actually do" (p. 35). North (2000b) notes that such kinds of scales only "give pictures of successive levels of language learning attainment" (p. 11). Fulcher (1987) also notes that the "present assessment scale [....] is attempting to describe not what actually happens in communicative situations, but what communicative theorists think happens in communicative situations" (p. 290). In brief, opponents of the theory-based approach claim that scales based on the assumption of an L2 developmental continuum are simply wishful thinking.

Second, the theory-based rating scales are also criticised for referring to the proficiency of native speakers when developing the top level of the scales. This is seen in many rating scales, and McNamara (1996) explains it as follows:

> This practice dates from the earliest oral proficiency interview test, the FSI Oral Proficiency Interview (OPI), where the highest level (5) is defined as follows: 'Native or Bilingual Proficiency: Speaking proficiency equivalent to that of an educated native speaker' and has survived (with some cosmetic modification) in many related rating scales (p. 184).

With rating scales for writing, the highest level is defined in the ASLPR (McNamara, 1996: 184) as, for example, "Written proficiency equivalent to that of a native speaker of the same socio-cultural variety". The TWE scale, the TEEP scale and the IELTS scale also refer to the proficiency of native speakers for the top level. As Alderson (1980) and Bachman (1990) point out, and a number of studies reviewed in Hamilton *et al.* (1993) and McNamara (1996) reveal, however, the level of native speakers varies dramatically depending on their educational and professional background, and native speakers did not get perfect scores but only relatively high ones on average. Furthermore, even if the level of native speakers would ideally be perfect, learners would be unlikely to achieve that level. Therefore, the reference to the native speaker for the top level on a scale is problematic.

Third, the theory-based rating scales have also been attacked in terms of validity. According to Upshur and Turner (1995), it is often the case that the descriptors of

such scales do not reflect the teaching objectives in a particular context. Consequently, the content of teaching in a classroom may not be compatible with the scoring standards. Additionally, it is often the case that the scales include descriptors relating to features which cannot be observed in the task that the students are being asked to do. For example, students are invited to do a speech task which does not specify the use of question forms, while descriptors of the scale may include features such as 'Fails to invert subject and verb in *wh*-questions'.

In the same vein as Upshur and Turner, McNamara (1996) points out the problem of validity with the scales. He asserts that the existing scales are not equipped with empirical evidence for their validity.

Matthews (1990) also criticises the validity of the rating scales. She summarises the problems in distinct areas: (1) some of the selected criteria in rating scales are found to be arbitrary and inconsistent; (2) there are descriptors of abilities within a rating scale that cannot be tapped by the tasks in a test to which the rating scale pertains; (3) the descriptors at times look ambiguous and are of little assistance to raters.

In research on the validity of an example of theory-based rating scales, the ACTFL guidelines, Fulcher (1996a) writes that those rating scales that include the ACTFL guidelines lack "empirical foundation" (p.164) and consequently evidence of validity, asserting that:

> It would seem fair to conclude that without a sound empirical basis for initial rating scale development, it makes little sense to investigate the validity of an oral rating scale *post hoc* [.....] Despite lack of empirical evidence, the model of language learning assumed by the ACTFL/ETS/ILR rating scales (often shortened to AEI) has become the basis for a whole approach to language teaching and testing in the US known as the Proficiency Movement (Higgs, 1984). The wide acceptance of the principles of the movement has led some authors to make a strong claim for the validity of the AEI rating scales, based on what is essentially face validity. Thus, for example, the most common defence for the validity of the AEI oral tests and rating scales is that of *experience* [emphasis added] (p. 164-5).

Finally, the theory-based rating scales, despite their name, do not reflect current learning theories. As Scarino (1995) points out, the basis for ordering descriptors as levels in the rating scales by the *a priori* approach is atheoretical. According to her, the descriptors are ordered very linearly, which ignores considerations of learning theory. She makes this claim, pointing out that the current rating scales using an *a*

*priori* approach are constructed too linearly to reflect the real learning process, which is characteristically spiral.

Critiques of rating scales using the *a priori* approach have been considered. All of these criticisms result from the fact that they are not based on actual performance data,[32] but rather on theoretical assumptions about developmental sequences. Given these problems, data-based approaches to constructing rating scales (e.g., Fulcher, 1996b; Upshur & Turner, 1995) have been suggested as a solution.

### 4.5.3 The data-based approach

As discussed in section 4.5.2, there has been much criticism of rating scales based on the *a priori* approach. In reaction to this, a data-based approach to rating scale construction has been proposed. For example, Griffin (1990), Alderson (1991), Fulcher (1993, 1996b) and Upshur and Turner (1995) constructed rating scales using this approach. However, although they all developed rating scales based on actual data, the methodologies they adopted are not identical.

First, Griffin (1990) produced a rating scale for reading assessment. Griffin's procedure for the construction of the scale is as follows: primary teachers attending a workshop got together to gather a range of performance indicators on the reading behaviour of primary students. The teachers were then asked to discuss whether the performance indicators properly represented students' reading behaviour, and to discard any that were unsuitable. The teachers were also guided to assess the indicators against students' reading behaviour in the classroom and to remove inappropriate ones from the pool of indicators. The sifted indicators were then calibrated by the IRT to decide the required attribute level of each indicator. As a result, each indicator could be mapped on the developmental continuum and grouped into seven bands. When doing field trials with the obtained bands before putting it into execution, Griffin (1990) stated that the ability level of a student could rarely be described using a single band. Put another way, a full description of a student's reading behaviour can only be achieved using several sorts of bands. Based on this position, a 0-3 point scale is used to describe the extent to which each band

[32] Brindley (1998) points out that although the ACTFL guidelines and the ASLPR claim to be empirically developed, there is no specific research evidence to support the claims.

represents a student's ability level. These are shown below:

> 3 If the student has established the behaviour pattern and consistently exhibits most of the indicators in the band, use a code of 3.
> 2 If the student is developing the behavioural pattern such that some but not all of the indicators for a band are often exhibited, use a code of 2 for that band.
> 1 If the student is beginning to show signs of the behaviour pattern of a band level in that only a little of the pattern is shown, use a code of 1 for that band.
> 0 If the student shows none of the behaviour pattern in a band use a code of 0 for that band.
>
> (Griffin, 1990: 298)

As a result, bands of reading ability from A to G, and a scale of the extent to which each band describes a student's ability, from 0 to 3, are used together in order to assess students' reading ability.

Second, Alderson (1991) constructed a rating scale by consulting the judgement of expert raters. Experienced raters for the ELTS writing tests were asked to choose sample scripts which could be representative for each band, before being guided to discuss them with other raters and decide the key features of each script. They then specified criteria for rating scripts on the basis of them. The criteria were then sorted in terms of levels or bands and were examined against sample scripts to ensure that they were free of anomalies and inconsistencies. This process of review was repeated.

Third, Fulcher (1993, 1996b) developed empirical rating scales for assessing fluency and accuracy in speaking tests.[33] The most salient characteristic of his scheme is that a scale is constructed through analysis, not through expert judgement, and that features obtained by his method can be described objectively.

As for a scale for fluency, Fulcher (1996b) analysed twenty-one transcripts of recorded interviews to define the phenomenon of fluency. After analysing the transcripts, he specified six observable speech phenomena that could potentially interrupt perceived fluency and influence a rater's judgement. These phenomena are as follows:

---

[33] Fulcher (1993) developed rating scales for two speaking qualities, fluency and accuracy, while Fulcher (1996b) presented one solely for fluency.

(1) Fillers such as 'er(m)'
(2) The repetition of the first syllable of a word or a full word
(3) The negotiation of reference indicated by the reselection of referring devices
(4) The reselection of lexical items
(5) Anacolouthon
(6) Longer pauses of three seconds or more

(Fulcher, 1996b: 215)

Of those phenomena, he focused on hesitation phenomena (including pauses) that non-native speakers exhibited in the test situation, and on their effect upon rating. He pointed out that when categorising hesitation phenomena it was the length and nature of pauses in speaking, not speed of speech and number of pauses that influenced the rater. To explain the nature of pauses, he created an eight-category coding system to code the phenomena from the transcripts. These are as follows:

(1) End-of-turn pauses: pauses indicating the end of a turn
(2) Content planning hesitation: pauses which appear to allow the student to plan the content of the next utterance
(3) Grammatical planning hesitation: pauses which appear to allow the student to plan the form of the next utterance
(4) Addition of examples, counter examples or response to support a point of view: these pauses are used as an oral parenthesis before adding extra information to an argument or point of view, or break up a list of examples
(5) Expressing lexical uncertainty: pauses which mark searching for a word or expression
(6) Grammatical and/or lexical repair: hesitation phenomena which appear to be associated with self-correction
(7) Expressing propositional uncertainty: hesitation phenomena which appear to mark uncertainty in the views which are being expressed
(8) Misunderstanding or breakdown in communication

(Fulcher, 1996b: 216-7)

He then described how each category was exhibited at various proficiency levels in the transcripts. It was revealed that oral rating scales could not be linear/monotonic, because students of a higher proficiency can commit more errors in the process of acquiring more language and experimenting with it. Therefore, the phenomena coded by the categories and the language ability were not always positively correlated. That is, it is not the case that a specific category is related with a specific language proficiency level. This is why he included almost all of the eight kinds of category within each band, when developing descriptors of each band using the categories.

The descriptors were applied to Bands 1-5, and he included two additional

bands (Bands 0 and 6) at the polar points of his scale, with Band 0 being described as 'less fluent than Band 1' and Band 6 as 'more fluent than Band 5'. The reason for this is twofold. First, Bands 1 and 5 cannot be absolute extremes of the language ability of the population. He thought that because the data obtained for this study could not represent a full range of language ability among the population, it would not be possible to describe anything less or more proficient beyond the data. Second, according to the hypothesis that raters usually avoid using the highest and lowest bands on the scale,[34] he intended to have raters use the full range of bands which were fully described.

Finally, Upshur and Turner (1995) suggested another possible method of constructing a rating scale. They point out that the rating scale they developed is an empirically derived, binary-choice and boundary-definition scale (EBB scale). To develop it, the six-member research team first chose a group of student performances. Upshur and Turner emphasise that every effort should be made for the selected samples to cover a full range of ability in the subject group. Each member of the researcher team was told to divide the samples into two halves, upper and lower, working individually rather than collectively. The team then got together to discuss their own dichotomous division and to adjust any differences. After agreement, they established a criterion in the form of a question that would allow them to divide the set of student performances into two groups. Each researcher was then guided to rank individually and impressionistically the three upper-half performances into the level of 6, 5 or 4, with at least one sample rated as 6. The whole team met to discuss and reconcile the results of ranking again. They used the reconciled rankings to formulate criterial questions, one to discriminate between level 6 performances and level 4 and 5 performances, and the other to discriminate between level 5 and level 4. They then repeated this procedure for lower-half performances. An example of a rating scale for grammatical accuracy derived from the procedure is shown in a figure below, reported here as Figure 4.1.

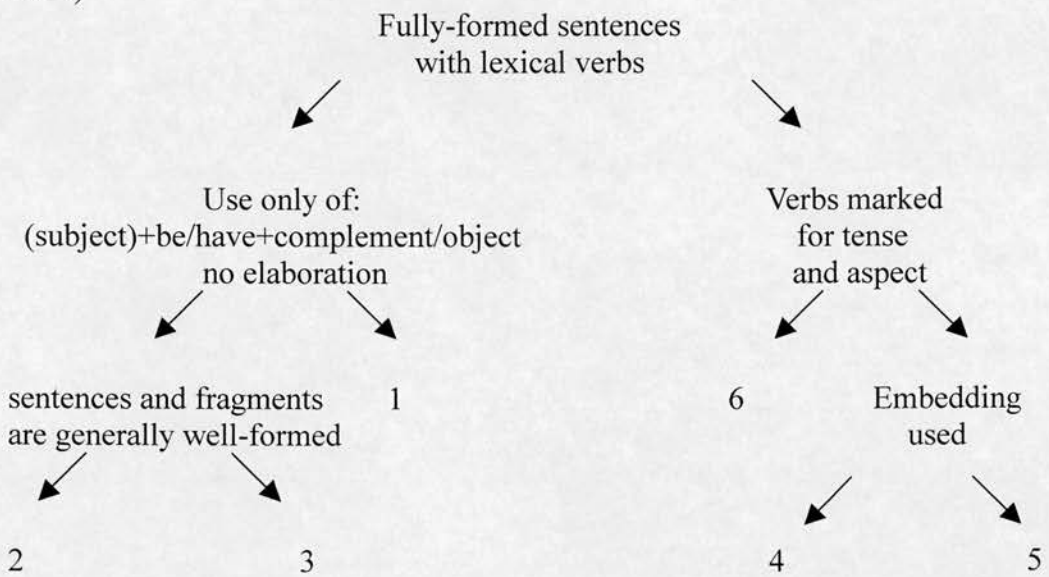
Upshur and Turner claim that their rating scale has three advantages over other scales: (1) it is easier to judge student performances as the boundary rather than mid-

---

[34] This phenomenon is referred to as 'central tendency' (McNamara, 1996). See section 6.2.3.2 for more discussion.

point of a level is formulated; (2) the reliability of rating is very high because the descriptors are so explicit; (3) as the scale is based on a local population and context, it has a positive effect on the validity of the rating.

Figure 4.1 Rating scale for grammatical accuracy (quoted from Upshur & Turner, 1995: 9)




[Originally Fig. 4]

As we have seen, methodologies for data-based approaches have been suggested by various researchers. These methods can be summarised as follows:

(1) In Griffin (1990), informants were asked to identify key target behaviours and descriptors were formulated from these. These descriptors were incorporated into a questionnaire to get teachers to rate the difficulty levels of each descriptor, and then the descriptors were calibrated using the Rasch model. With the calibrated descriptors, a rating scale was formulated with cut-off points on the scale;

(2) In Alderson (1991), a rating scale was developed using raters' intuitive identification of key 'features' at different levels with regard to performance samples ranked in a consensus order;

(3) In Fulcher (1993, 1996b), a rating scale was developed through the analysis of occurrences of key performance features in samples. Here experts' judgement was not employed;

(4) In Upshur and Turner (1995) a rating scale was developed through raters' comparison of pairs of performances, stating which was better and why, in order to identify the salient features. A yes/no binary algorithm was therefore constructed;

As with the *a priori* approach, there are criticisms of the data-based rating scales. Turner and Upshur (2002) summarise these criticisms in three points. First, the development process needs a great deal of time. Second, since the development is based on a limited sample, the scales lack general applicability. Third, basing the scale on data rather than theory means that the data-based scales are atheoretical. Chalhoub-Deville (1997), quoted in Turner and Upshur (2002) responds to the last two points as follows:

> empirically based scales developed for one task type are indeed theoretical, and represent subtheories of a more general theory. In other words, due to their specific context, they represent a particular instance of a more global language proficiency theory. The lack of generality of these rating scales is not in dispute, but more general, theory-based rating scales have not been shown to be equally valid for the various task types that empirically derived scales are designed for (p. 52-3).

I find this convincing, and believe that a theory does not have to cover a whole span of learning from zero to perfection, nor it is possible to deal with every context where language is used. In light of this, even though a data-based rating scale may be limited to the data employed, it is nevertheless worth constructing. However, I concede that the great deal of time required to construct this type of scale is its primary drawback.

### 4.5.4 Conclusions

In this section, two main approaches to rating scale construction have been examined: the *a priori* approach and the data-based approach. The former is based on theory and theorists' assumptions, whilst the latter is based on real performance data. Consequently, the former may be inappropriate for particular contexts and learners because it is not context/learner-specific, despite looking neat and logical; whilst the latter is concrete, despite looking messy. In this way, these two approaches are the exact reverse of each other in their merits and demerits as well as in their methods of constructing a rating scale. It is not an easy matter to decide which of the two is

better. Given the current great concern for validity in assessment, however, in my opinion the latter would be better than the former, which may be criticised for its invalidity. Nevertheless, the time, effort and resources required to gather and analyse data for the data-based approach can be an obstacle.

## 4.6 The users of rating scales: raters

### 4.6.1 Introduction

Even if a well-developed rating scale is available, how well it is used depends on another factor: the rater. A scale does not function on its own, but needs to *be used* by a mediator, the rater. The role of raters cannot be exaggerated, because they are at the centre of the rating process, and the use of the rating scale depends on how they interpret and apply it. Lumley (2002) also addresses the importance of their role as follows:

> [The rater decides] which features of the scale to pay attention to; how to arbitrate between the inevitable conflicts in scale wordings; and how to justify [his/]her impression of the text in terms of the institutional requirements represented by the scale and rater training (p. 267).

While confirming the importance of raters, this comment implies that they are subjective even when using a rating scale. This is partly because, as noted in section 3.3.2, they bring their own background to the rating process (Weigle, 2002), and this affects the use of a rating scale and rating. The effects of their background on rating have been studied in order to promote reliable and valid rating. The types of background dealt with in the literature are vocational (whether the rater has a job related to linguistics, English as a native language teaching or ESL teaching), nativeness (whether the rater is a native speaker of the target language) and personality (whether there is any relationship between the rater's personality and the ratings he/she gives to writing). I will consider each of these aspects in detail in section 4.6.2. In section 4.6.3 I will discuss the effect of rater training, which may be a means of lessening the impact of these variables on rating.

### 4.6.2 The effect of rater background on rating

First, with regard to the issue of the rater's vocational background, the main concern

has been whether only a rater with a linguistics-related job would be suitable as a rater, or whether people in fields not related to linguistics can also act as raters (or whether people in the field related to linguistics can also be raters for a specific occupation-related test[35]). Many studies have been conducted, and their common conclusion is that linguistically-trained raters (such as ESL faculty) and non-linguistically trained raters award similar overall grades, but give different ratings for some assessment categories. Researchers explained this discrepancy as resulting from the difference between the raters' attention to or concern for assessment, and their perception of the rating scales employed.[36]

Brown (1991) investigated the rating of two groups of English faculty and ESL faculty. The raters were asked to rate holistically a total of one hundred and twelve scripts on two different topics, half of them written by ESL students and the other half by native speakers of English. They were also asked to identify the best and worst features of each script in terms of Cohesion, Content, Mechanics, Organisation, Syntax or Vocabulary. There were no significant differences in mean score between the two test-taker groups or in the ratings given by the two rater groups. Furthermore, both rater groups most often identified Content as the best feature. However, there were differences between the groups. For the best feature, the English faculty group paid more attention to Cohesion and Syntax than the ESL faculty group, whereas the ESL rater group was more concerned with Organisation than the English faculty group. For the worst feature, both groups selected Syntax most often. The English faculty paid more attention to Mechanics than the ESL faculty group, whilst the ESL faculty group was more interested in Content than the English faculty group. These differences indicate that even though raters arrive at the same score, they do so from

---

[35] Lumley (1995) investigated this issue for an occupation-related test, a test of English for health professionals. His study examined whether, in this performance test, language-trained raters could be employed instead of doctors, who are the occupational raters. The ten ESL raters and ten doctors in this study were asked to assess the candidates' overall communicative effectiveness in their performance, and considerable agreement emerged between these two groups, to the extent that he concludes that ESL raters can quite reasonably be expected to make judgements in this kind of occupational test.k

[36] There are also studies where the difference between rater groups is reported to be due to factors other than differences in the raters' concerns and perception of rating schemes. Song and Caruso (1996) found this difference between them to be due to the scoring method employed. In their study, it was found that English faculty and ESL faculty scored significantly differently in holistic scoring but not in analytic scoring. On the other hand, Michael *et al.* (1980) ascribed this difference to the different prompts in the scripts given to the two groups of raters (professors group in English department vs. professors group in other disciplines), rather than any difference in rater background.

different perspectives depending on their background.

Brown (1995) also explored this issue, with Japanese language as the target language. Native and near-native speakers of Japanese, who were either teaching Japanese as a foreign language or tour guiding in Japanese, used two rating schemes to assess fifty-one spoken performances on video. One of these was related to linguistic skill (Grammar and expression, Vocabulary, Fluency, Pronunciation, Vocabulary and Use of polite forms), and the other to task fulfilment (Enthusiasm, Empathy, The ability to make something sound interesting, Persuasiveness, Awareness of the interlocutor's needs/wants and so on). Their ratings using these two rating schemes were compared. Whilst they behaved similarly to each other in ranking the candidates and in severity of rating, they differed in perceiving particular features of language and task. The Teachers group tended to be more severe on linguistic aspects (Grammar and Expression, Vocabulary and Fluency), and the Guides group overall tended to be more lenient. They also differed in the use of the rating scale: the Teachers group was more reluctant to award extremely low and high scores than the Guides group.

O'Loughlin (1992) compared the ratings of teachers of English as a native language and teachers of English as a second language (hereafter referred to as English group and ESL group respectively). Four hundred and eighty-four written samples were gathered from secondary school students sitting the English test designed by the University of Melbourne to select students for undergraduate courses. Half of these students were native English speakers and the other half were non-native English speakers. After a training session the raters rated these scripts using both a holistic scale and an analytic scale. The ratings showed that there was greater agreement between the two groups in the holistic scoring than in the results combining holistic scoring and analytic scoring which means that they differed in terms of analytic scoring. The two groups differed in severity. That is, the English teachers group was more severe in analytic scoring than the ESL teachers group. Furthermore, the English group was influenced by all the categories in the analytic scale when making their global judgements on both the native- and non-native speakers's essays, whilst the ESL group was more influenced by Content and Organisation than the other categories for both groups of students, and by Grammar

and cohesion for only the non-native speakers' essays. Therefore, O'Loughlin concludes that the two rater groups behaved differently, and that holistic scoring can "mask important differences between raters" (p. 39).

Second, with regard to nativeness, there is substantial research comparing native speakers and non-native speakers of the target language. The foci of these studies are generally on differences in the severity of rating and differences in the use of the rating scale. Hill (1997) compared the rating of native (Australians) and non-native speakers (Indonesians) in the writing section of the English Proficiency Test for Indonesia (EPTI), which was developed to assess English proficiency as relevant to classroom teachers. This comprised reading and writing sections. The raters were asked to assess the writing subjectively. During the rater training session before the scoring session, they were expressly instructed not to refer to the proficiency level of ideal native English speakers for the top level, but rather to think of what would constitute good writing for teaching English writing in Indonesian high schools, given that the local variety of English is the criterion. She compared the rating pattern between the two groups in terms of consistency, severity and use of assessment criteria. Although the native speakers group were more experienced raters than the non-native speaker group, it was found that all but one of the non-native speakers rated consistently. Furthermore, the non-native speaker group agreed more closely with each other regarding severity than the native speakers group. Additionally, regarding the use of assessment criteria, both groups made Overall Impression and Control of Linguistic Features the categories in which it was hardest and second hardest to gain a high score. There was, however, a difference between the two groups in the use of the Coherence and Cohesion category: first, the native speaker group rated more severely at the cut-off point (pass mark) than the non-native speaker group, so fewer candidates were rated above this point by native speakers than by non-native speakers. Second, the non-native speaker group were more reluctant to assign the top level of the scale than the native speaker group. She suggested that this was because the non-native speakers had still applied a native speaker standard for the top level even though they were trained not to do so. From these results, she concluded that non-native speakers were not less suitable as raters than native speakers, at least on the local English test.

Shi (2001) investigated the differences between native- and non-native speaking EFL teachers in judging Chinese students' English writing. Twenty-four non-native speaking EFL teachers and twenty-four native speaking EFL teachers were asked to rate subjectively and holistically ten written samples by third-year students in the English department of a university in China. To find out whether these raters made different qualitative judgements, they were also asked to write a self-report giving both three reasons and comments for their judgements in their order of importance for their rating. These reasons and comments were usually short phrases, but for the purpose of analysis they were coded as key words in the area of L2 writing or raters' attitude to the writing. The former included categories such as Thesis, Argument, Ideas, Content, Logic, Organisation, Grammar, Vocabulary, Paragraph, Clarity, and Support, whilst the latter included categories such as Good, Poor, Clear, Unclear, Balanced and Unbalanced. Analysis of the ratings and the raters' self-reports showed that, although the two groups gave the writing samples similar scores, they differed in the frequency of types of comments or criteria for their judgements. The native speaker group attended more positively to Content and language, whereas the non-native speaker group were more negatively concerned with Organisation and length (both groups attended positively to Content and negatively to Intelligibility).

Third, with regard to personality, research has been conducted mainly using the Myers-Briggs Type Indicator (MBTI)[37] to classify raters' personality type. Carrell (1995) describes many studies on this, including those by Gowen (1984) and Jensen and DiTiberio (1989), which attempted to find any kind of relationship between a rater's personality type and ratings. Carrell (1995) was more comprehensive than these studies in that she attempted to find not just these relationships, but also relationships between the writer's personality type and the ratings, and between the rater's personality and the writer's personality. To this end, she asked twenty English composition instructors at a university to rate forty-three writers' scripts using a slightly modified version of the TWE scoring guide. She also had all the participants take the MTBI. The writer's personality type was found to have a significant effect

[37] The MBTI (Myers, 1962, 1987) is an inventory of self-report types for the measurement of personality type in a variety of settings (educational, career, and family counselling). It consists of ninety-four items intended to identify an individual's basic preferences. These items are classified according to the four bipolar scales: Extroversion-Introversion, Sensing-Intuition, Thinking-Feeling and Judging-Perceiving (Brown, 1987; Carrell, 1995).

on the ratings for the bipolar scale of Extroversion-Introversion. The writing of Introverted writers was rated more highly than that of Extroverted writers. Additionally, the rater's personality type was also found to affect significantly the ratings, this time on both the bipolar scales of Sensing-Intuition and Thinking-Feeling. Raters who were Sensing or Feeling rated the essays more highly than those who were Intuitive or Thinking, respectively. However, there was no significant relationship between the personality types of writers and raters: for example, Introverted raters did not tend to rate scripts of Introverted writers better or worse than those of Extroverted writers.

The background features of raters discussed above have thus been revealed to have an effect on rating. This finding appears to be helpful in understanding and interpreting a rater's bias (Kondo-Brown, 2002) in severity and in the assessment categories. It seems that we will have to accept that all candidate raters have this bias and thus vary in their ratings because they have various backgrounds.[38] The fundamental concern is not just to understand and describe this bias in rating, but to find ways to improve the quality of rating. Therefore, there is a need for various methods for these, such as rater training and multiple scoring. This issue will be discussed in the next section.

### 4.6.3 The effect of rater training on rating

As mentioned in sections 3.2 and 3.3.1.1, early research on direct writing assessment mainly focused on the reliability of scoring, especially inter-rater reliability, and paid less attention to test validity. This was in order to show that the performance of test-

---

[38] One option for this variable could be to use an automated/electronic rating system (i.e., an e-rater). Warschauer and Ware (2006) reviewed various automated writing evaluation systems, such as Project Essay Grade, the Writer's Workbench, My Access!, Criterion and the Intelligent Essay Assessor, which were all designed to provide scores and feedback on writing, and have been developed, investigated and used since the 1960s. The advantage of these systems is that they can provide instant, individualised assessment and feedback on a wide or limited range of assessments and there are also many studies to show their effect on improving learners' writing. However, their usefulness still appears to be questionable, as Warschauer and Ware (2006) mention, for the following reasons: the studies have usually been carried out by the companies that produced the systems; few of the studies have been published in journals; and such systems may "[draw] students away from composition that is purposeful, audience focused, and expressive toward that which is formulaic and stilted" (p. 175). It may also lead them to focus on the writing features "most easily detected" (p. 170) by the scoring softwares, such as spelling, capitalisation, word form, discourse elements (such as inclusion of background, conclusion, main points and supporting ideas) and length, rather than a sense of audience, content and style.

takers could be rated fairly and consistently (Weigle, 1994). I would suggest that it was also intended to convince stakeholders such as test-takers, their parents, teachers and users of the ratings of the value and acceptability of direct tests. Various methods such as multiple scoring and rater training procedures were adopted to achieve acceptable reliability. I will focus on rater training here, which has been explored by many researchers because of its controversial effects.

Rater training usually takes the form of a session where raters are introduced to an assessment criterion, asked to rate some performance samples using the criterion, and then to compare and discuss their ratings in order to help them achieve a common interpretation of the criterion.

Studies on the effects of rater training have focused on rater consensus and validity. Some of these studies confirm the positive effect of rater training. Freedman (1981) found that training did change rater behaviour, given that the raters who discussed the meaning of a topic in some detail generally gave higher scores.

Weigle (1994, 1998) analysed her data both qualitatively and quantitatively. In her first study she asked sixteen inexperienced raters to score the same essays before and after rater training, and to complete a verbal protocol during the scoring. What she found from the verbal protocol analysis was that rater training helped the raters understand and apply the rating criteria as intended, modify their expectations of the characteristics of the test-takers, and raise their awareness of their fellow raters. Inter-rater agreement was not found to be an overriding concern amongst the raters, even though this is normally assumed to be one of the purposes of rater training.

In her quantitative study of the same data, Weigle (1998) also confirmed the positive effect of rater training. It was found in this study that rater training helped to improve internal consistency and train extreme raters who rated most severely. However, it is noteworthy that even though the range of severity among the raters decreased, there were still substantial differences in severity between them, which implies that avoiding differences between raters cannot be a realistic aim of rater training.

Other studies, on the other hand, verify that rater training is not as effective as intended. Brown (1995), who investigated the rater variable on the Japanese Test for Tour Guides, found that rater training did not help raters understand rating scales as

intended. Although the raters in her study, who were both native and near-native speakers of Japanese, attended a one-day rater training session, there were still differences between the two groups in tolerance to breakdowns in particular features of communication in the spoken performance on video. According to her this has significant implications for construct validity:

> If each group were to develop its own assessment framework, [.....] they may, in fact, through the inclusion or weighting of specific criteria, produce schemes which led to quite different evaluations of candidates' ability. This then raises the issue of whether one view is necessarily more valid than another: who should be devising assessment schemes relating to occupational proficiency? (p. 13).

Kondo-Brown's (2002) study also revealed that rater training was not very effective. The three raters who took part in the study were asked to assess essay samples written in Japanese by native English speakers whose Japanese proficiency levels varied. For scoring, the raters were provided with a modified version of Jacobs *et al.*'s (1981) rating scale adapted to the characteristics of Japanese language. The raters were trained for a total of three hours for inter-rater reliability before and between scoring. They were asked to assess ten scripts using the scale and to discuss ways of minimising rating discrepancies. As a result, although their ratings were highly self-consistent (i.e. intra-rater reliable), small but significant discrepancies between the raters were found in overall severity. This confirms that a rater's tendency towards overall severity or leniency in performance assessment still persists after rater training.

Lumley (1995) also emphasises that, even after training, differences between raters remain. Therefore, he is sceptical of rater training for the purpose of decreasing differences among raters. Rather, he suggests the need for rater retraining. Considering the inconsistent rating of the very experienced ESL raters in his study, who had a long history of reliable rating on the performance test, he suggests that regular retraining could be helpful for all raters. This view is also stated by Lumley and McNamara (1995).

From the studies discussed above, it appears that rater training is both effective and ineffective. "Rater training can reduce, but cannot easily eliminate a rater's tendency for overall severity or leniency in judging performance.... [Rather], rater training is successful in making raters more *self-consistent*" (emphasis in the

original) (Kondo-Brown, 2002: 4).

The case for rater training is supported by its effectiveness in improving self-consistency, namely, intra-rater reliability. However, when it takes the form of training sessions as discussed above, it needs time, budgetary and human resources. Therefore, as an alternative to this, self-training has been suggested. Kenyon (1997) studied the possibility of rater self-training, and concluded that rater self-training can be used as an alternative to group rater training.

As discussed so far, rater training and rater self-training are controversial. This is partly because it is not always effective, and partly because it leads raters to ignore their experience and expertise as raters for the sake of rater consensus. On the first point, we may conclude that although rater training cannot eliminate the differences in severity and leniency between raters, it is meaningful in that it can help give raters a shared understanding of the rating scale employed, and help them be self-consistent. On the second point, I believe that rater training can be defended as follows:

> ...essay-scoring and rater-training procedures are presumably founded on the premise that an essay examination is measuring a particular ability which can be defined operationally and measured accurately if raters can be trained to agree on the definition of the ability. From this point of view, it is essential for raters to put aside their own subjective experience in order to adopt the agreed upon scoring criteria for the examination (Weigle, 1998: 264).

In sum, rater training is meaningful and worthwhile in that it helps raters self-consistently understand and measure a specific construct which is intended to be measured in a test or a rating scale. This viewpoint allows for some variability in the rater's backgrounds and in their reactions to a script, since it does not emphasise rater consensus. However, as Weigle (1998) notes, too much emphasis on intra-rater consistency at the expense of inter-rater agreement is not desirable, because it could lead to deviation from construct validity. When inter-rater reliability is discarded and each rater is allowed to interpret the rating scale in an idiosyncratic way, only paying attention to internal consistency, the construct which they assess could be different from the one intended by the test or rating scale developers, even though the raters might be internally consistent. Therefore, there should be a balance between these considerations on the part of rater trainers and raters.

## 4.7 Conclusions

I have discussed various aspects of rating scales, covering their history, nature, type, development method and the users of rating scales, and have shown that all kinds of rating scales have certain limitations, whether they are holistic, analytic, primary-trait, theory-based or data-based. Hwang (1930) summarises the limitations of rating scales as follows:

> The effective use of measuring instruments again depends upon a frank recognition of their limitations. This is especially true in the field of rating. [.....] we come across many sources of errors and unreliability. A portion of those errors, no doubt, comes from scales themselves [...] (p. 1).

However, they have positive effects as well, in terms of reliability and validity. According to Webb (1915) quoted in Hwang (1930), there are two kinds of errors: systematic errors and random errors. Systematic errors have a consistent pattern, deviating from some average scores in the same direction and by similar amounts. Random errors do not show such uniformity in the amount of deviation in all the scores of a given group. Therefore, although their adequacy in terms of inter-rater reliability might be questionable, rating scales could help reduce random and systematic errors within a rater's rating, as revealed in Hwang (1930). This is a convincing argument for their use.

In Chapter One I argued the need for rating scale development. In Chapters Two to Four, I reviewed three main issues in the development of rating scales: the writing ability that is to be assessed, the writing assessment task in which the rating scale is to be applied, and the scoring criteria to be used for assessment. In Part B, on the basis of these considerations, I will describe the development of a rating scale for the English Writing course in FLHSK.

# PART B.

# THE STUDY

# CHAPTER FIVE. METHODOLOGY

## 5.1 Introduction

This is the first chapter of Part B, which will deal with the development of the rating scale as a proposed solution to the problems raised in Chapter One, informed by the studies mentioned in Chapters Two to Four.

I will start with the preliminary questionnaire survey, which was carried out before the main study. This was one of three methods I used to empirically investigate the arguments made in Chapter One regarding why a rating scale might need to be developed for Korean students (the other two methods were to investigate subjective holistic scoring and the FCE scoring which will be discussed in Chapter Six).

After discussion of this survey in section 5.2, I will describe the main features of the methodology of the study, such as subjects, writing tasks, raters and research design. Given that the rating scale was to be developed by a data-based approach, I first needed data, that is, writing samples. With regard to this, I will discuss who wrote the writing samples (i.e., subjects) in section 5.3, which tasks the subjects were asked to do (i.e., writing tasks) in section 5.5, who was asked to assess the obtained writing samples according to various kinds of scoring schemes for this study (i.e., raters) in section 5.4, and how this research was designed and conducted in section 5.6.

## 5.2 Preliminary questionnaire survey

### 5.2.1 Introduction

Before entering into the research on rating scale development for writing assessment in FLHSK, a preliminary survey was conducted on English teachers at FLHSK by means of a questionnaire (which will henceforth be referred to as Questionnaire I, and can be found in Appendix 2). In Chapter One, I argued on theoretical grounds that there was a need to develop a rating scale for the English Writing course at FLHSK. However, this was only a personal opinion based on previous studies and professional judgement. So, before beginning to develop a rating scale, my opinion needed to be examined through empirical research. To this end, I used three methods.

One was to ask teachers at FLHSK whether or not they thought there was a need to develop a rating scale for classroom writing tests in Korean high schools; the other two methods involved finding out whether the rating schemes available at present were satisfactory. For the second method I intended to ask English teachers to assess students' scripts according to their own subjective judgement, and for the third I intended to ask them to assess the scripts using one of the published rating scales. In this chapter, I will discuss the first method, i.e., the questionnaire survey. The second and third of these three methods will be discussed in Chapter Six.

### 5.2.2 Purpose

The purpose was mainly to find out how English teachers in FLHSK assessed the writing of their students, and whether they thought it desirable to develop a rating scale for the students.

### 5.2.3 Respondents

The respondents to Questionnaire I were one hundred and nine English teachers in fourteen FLHSK.

### 5.2.4 Procedure

The questions in Questionnaire I were developed on the basis of Cohen, Manion and Morrison (2000), Converse and Presser (1986) and Gillham (2000), as well as informal talks with three English teachers about the current situation regarding writing assessment in schools. The questionnaire was written in Korean so as to help respondents to feel comfortable.

I decided to carry out the survey using Questionnaire I on-line. To this end, the email addresses of teachers at fourteen schools (of the twenty FLHSK at the time) were obtained either by browsing the school websites or contacting acquaintances in the schools.

Before starting the survey, an introductory email was sent informing teachers that they would receive a questionnaire for research purposes in a few days. The questionnaires were posted on the 21st of May 2003 by a research agency on the Internet (www.research.joongang.com). Some of the English teachers responded very

quickly, whereas others did not. Consequently, the questionnaires were posted again on the 27th of May 2003 to those who had not yet replied. As a result, more responses were obtained. The survey finished on the 31st of May 2003.

### 5.2.5 Results and discussion

Out of the one hundred and nine English teachers in the fourteen schools, forty-two (38.5%) filled in the questionnaire. However, seven of the forty-two questionnaires were not counted because respondents did not follow the instructions attached to them. All in all, thirty-five respondents were counted for analysis. Non-returns may have various explanations, including: (1) some of the teachers' mailboxes may have been full, so that they did not receive the questionnaire; (2) some of them might not use the email address listed on the school website; (3) some of them may have set up a device for blocking emails from unknown senders and/or spammers; (4) some who received the questionnaire may have lacked the time or interest, or motivation to answer.

It is noteworthy that none of the thirty-five respondents actually reported that they taught the English Writing course. However, this does not mean that the course was not being taught. Teachers were teaching two courses at the time of the survey but were allowed to report only one, so they might have preferred to report the major course such as English I, English II and English Reading rather than minor courses such as English Writing, English Grammar, English Listening and English Culture (see the results on Q7 in Appendix 3). I found this to be the case from talks with several respondents on the phone after the survey closed. Furthermore, from the responses to Q9 for respondents who did not carry out writing assessment in their own courses, five of the twelve respondents did not have writing assessment sessions in their own courses because there is a separate course for writing. Therefore, we can safely conclude that the English Writing course is currently being taught in FLHSK.

Another point to note is that, with regard to subjective holistic scoring, eleven (31%) of the twenty-three respondents who carried out writing assessment in their courses reported that they assessed according to their subjective judgement without using a rating scale, and that they were worried about its validity as well as reliability. It was also found that eight (23%) of the respondents did develop rating scales for

themselves or with other colleagues. In that case, however, it was found that they were not satisfied with them, as they were worried about the reliability and validity of scoring.

One of other noteworthy findings is that thirty-two teachers (91%) responded negatively to the questions asking whether any rating scales had been suggested by organisations or other societies (Qs 10 and 11). Also, it was found that thirty of the respondents (85%) thought it desirable to develop a rating scale specifically for Korean high school students (see the responses to Qs 23 and 25 in Appendix 3). This provided empirical evidence that English teachers agree that there is a need to develop a rating scale for the English Writing course in FLHSK. This meant that I could start the study on rating scale development.

Having asked English teachers about the current situation in writing assessment and their opinion on the need for a new rating scale to be developed for Korean students, I embarked on the main study. First, I obtained writing samples to be used as data for a data-based rating scale development. These were then assessed by raters using a subjective holistic scoring scheme in order to investigate any problems and tendencies with the scoring scheme and obtain the raters' ratings of the samples, which would be used to group the samples into six bands before the rating scale was developed. I also had the samples assessed by one of existing rating scales to find any problems with it empirically, and analysed the writing samples according to a coding scheme in order to develop a rating scale. I will begin by discussing the features of this study, such as subjects, raters, writing tasks, and then deal with the research design in the following sections.

### 5.3 Subjects

For this study, three hundred and thirty-three students were asked to complete two kinds of writing tasks for a pilot study and two kinds of tasks for the main study with regard to obtaining writing samples that were needed for the development of a rating scale. They consisted of one hundred and seventy 1$^{st}$ year students and one hundred and sixty-three 2$^{nd}$ year students at Kwacheon foreign language high school near Seoul, Korea. They were from eight classes with about forty in each class, were aged

between fifteen and seventeen, and were about half male and half female. I was able to have their writing thanks to three teachers who agreed to collect the scripts from their students (henceforth known as T1, T2 and T3).

Since they attended a foreign language high school, it can be assumed that the average level of their English was above that of students at a general academic high school (see section 1.6.1).

In the Kwacheon foreign language high school, the English Writing course was introduced to 3rd year students. This meant that the subjects had not taken it yet, but they often had opportunities to write essays as part of the performance assessment of other courses during their 1st or 2nd year.

### 5.4 Raters

I needed as many raters as possible for this study, whom I would ask to assess students' scripts, keep a diary and do a think-aloud. However, it was very difficult to obtain enough teachers for these jobs. Teachers at high schools in Korea have a considerable workload, and few of them were willing to spare the time for extra tasks such as participating in research. Fortunately, one English teacher who worked at Kwacheon foreign language high school agreed to be one of the raters (he participated in this study as both T1 and Rater A). For the other raters, it happened several times that teachers committed themselves but then withdrew. I tried to find teachers for this work by advertising on the Internet to English teachers at the fourteen FLHSK and contacting acquaintances. Finally, two teachers from different general academic high schools helped me out and carried out the rating. Although it would have been the best if they had worked in foreign language high schools, ultimately this was not possible.

Thus I finally obtained three raters, whose profiles were as follows:

Rater A

- male;
- worked at Kwacheon foreign language high school and taught four first-year classes of the eight classes in the subject group (i.e., T1);
- taught the courses for English Grammar and English I;

- had twelve-years teaching experience in middle or high schools;
- had obtained a Master degree in the US and has completed a PhD coursework in Applied Linguistics in Korea;
- carried out writing assessment three times per semester in his own normal class;
- assessed the scripts from the subjects in one first-year class and one second-year class (henceforth known as 'Group A').

Rater B

- female;
- worked at Cheongwon general academic high school;
- taught the course for English I;
- had five-years teaching experience;
- had obtained a Master degree in Applied Linguistics in Korea;
- carried out writing assessment about once per semester in her class;
- assessed the scripts from the subjects in one first-year class and one second-year class (henceforth known as 'Group B').

Rater C

- male;
- vice-principal at Soongmoon general academic high school;
- taught the general English course at the school;
- had thirty-four-years teaching experience;
- had obtained a Master degree in English Literature;
- carried out writing assessment once or twice per semester in his class;
- assessed the scripts from the subjects in one second-year class (henceforth known as 'Group C').

For their assessment, keeping a diary and doing a think-aloud, Raters A, B and C were paid a fee.

### 5.5 Writing tasks

The writing tasks were produced by the researcher specifically for this study. Writing tasks can be controlled, guided or free writing. Given the purpose of the English Writing course at FLHSK, free writing tasks were chosen for this study.

I needed to decide how to set the task in terms of genre, topic and other rhetorical specifications. Reviewing the literature discussed in Chapter Two, the context of the teaching and assessing of writing in Korea, the aim of the course objects and the textbooks for the course, I tentatively decided on the appropriate writing tasks. Before deciding to employ the tasks for the study, however, I piloted them in all of the subjects. I will now discuss this pilot study to decide the tasks for the main study.

### 5.5.1 Pilot study for writing tasks

It was important that the subjects demonstrated their writing ability, so in order to ensure that the writing tasks for the study were appropriate for this, it was desirable to conduct a pilot study. The outcome of the pilot study would help determine the difficulty and appropriateness of the task. If the tasks were inappropriate or too difficult for the subjects, they would not allow the subjects to show their writing ability, and it would not have been meaningful to develop a rating scale based on such writing data.

To begin with, the types of task had to be decided. Ideally, for the purpose of this study the students would be asked to do a variety of tasks, such as business or casual letters, diary entries, invitation letters and essays. However, because of the practical constraints, I had to narrow down the number of tasks and choose the most appropriate ones. Supposing that all the writing tasks could be divided into two main categories of formal and informal writing, I decided to use both of these for this research and select a representative/common one of each type. My aim was to provide balanced writing samples from both informal and formal types of writing tasks, and in turn make the First version of the new Rating Scale (RS1) applicable to the assessment of both styles.

An informal letter was considered appropriate for the genre of informal writing, and a formal essay for the genre of formal writing, as these were genres for which students had submitted writing tests and assignments, or which were included in

every writing textbook for schools in Korea.

Once the genres were decided, the topics for the tasks had to be selected. They were chosen from three textbooks published for the English Writing course. They were as follows:

- your life at school;
- the advantages and disadvantages of living in a big city

After the genre and topic for the writing tasks were decided, prompts needed to be written. They were designed to have the following characteristics.

First, they were presented in the form of complete sentences including genre and audience, etc., rather than titles such as "school life" or "advantages and disadvantages of living in a big city". It was expected that this would help the subjects understand more clearly what they were expected to do and help them work straight towards the goal. In addition, all prompts were given in Korean rather than in English, to avoid the effects, if any, of a reading ability variable on performance.

Second, several key points on content that subjects could optionally use in the writing tasks were presented in bullet points as part of the prompts. This was to reduce the effort and time subjects needed to think about content *per se*, and allow them to spend the given time writing rather than thinking about what to write, in accordance with Read's (1990) recommendation that students should be provided with all the relevant content material within the prompts (see section 3.3.2 for more details).

Third, the subjects were given one writing task for each writing session. They were not given choices. If test-takers had been allowed to choose their tasks, it would have been desirable to measure whether the tasks were equally difficult in order to assess the subjects fairly, but this is not easy (Weigle, 2002). Jacobs *et al.* (1981) take the same view, claiming that "allowing a choice of topics introduces too much uncontrolled variance into the test" (p. 16). Therefore, to ensure that observed differences in scores are mainly due to real differences in writing competence, not to the different topics or tasks, all subjects were asked to do the same task during each writing session for this study.

Finally, the minimum length of work – at least two hundred words – was specified in the prompts because it is believed that when subjects are informed of the minimum length, they are likely to be able to infer how extensive the content of their writing should be (Weigle, 2002). The length of one hundred words is commonly considered the minimum for continuous writing. Weir (1993) claims that the scripts that test-takers are required to produce have to be long enough to be marked reliably. Jacobs *et al.* (1981) also found that in a thirty-minute composition test, students could write about a page or more. Hence it was considered appropriate that the subjects should be invited to write an essay or letter of at least of two hundred words during a fifty-minute class-session.

In brief, the prompts were made as specific and purposeful as possible in order to avoid as much as possible variable-effects that might affect the subjects' performance and fair assessment of their work. As a result, two tasks for a pilot study on tasks were devised, Tasks 1 and 2, and the resultant prompts for each task were as follows:

- For Task 1

Question: **Write a letter informing a foreign friend about your life at school**
You may include an introduction to the following places:
General characteristics of high schools in Korea;
Daily life at schools;
Some courses that are taught at school or interesting to learn
After-school activies/clubs
Your writing should be at least 200 words long and must be completed within the class session.

- For Task 2

Question: **Write a formal essay explaining the merits and demerits of living in a big city to your teacher**
You may include the following features as merits:
Convenient transportation;
Rich educational and cultural facilities;
Many opportunities for various jobs and experiences.
You may include the following features as demerits:
Air or noise pollution;
Lack of solidarity among neighbours.
Your writing should be at least 200 words long and must be completed within the class session.

All the subjects carried out these tasks over two sessions in T1, T2 and T3's own normal classes in July 2003. After they completed the tasks, to investigate whether the tasks had flaws or failed to help draw out the subjects' writing ability, I considered it desirable to ask the subjects how they found the tasks after they completed them, and then revise them for the main study. The investigation was carried out through another questionnaire, henceforth referred to as Questionnaire II (see Appendix 4). Questionnaire II was answered by them, and their answers were coded and analysed for frequency using the SPSS computer package.

As for Task 1, 15.1% and 50.6% of the students found it either fairly or slightly difficult respectively[39]. Of the respondents who answered this, 27.6% had difficulty with the genre format of the task, 50.4% had problems with the content, and 48.6% had difficulties with the style/register of the task. However, given that 90.8% of them answered that they had no difficulty in fully understanding the prompt, the prompt *per se* was not hard to understand. Almost all of the respondents (74.4%) answered that it was useful to have key points on the content of their writing, and most (68.7%) found it useful that the audience was specified in the prompt. As a result, many of them (66.9%) answered that they had been able to demonstrate their writing ability through Task 1.

As for Task 2, 21.7% and 47% of the respondents found it either fairly or slightly difficult, whilst 27.4% of the respondents found it either reasonably or very easy. Of the respondents who answered fairly or slightly difficult, 31.9% had difficulty with the genre format of the task, 45.8% had problems with the content of the task, and 40.6% had difficulties with the style/register of the task. 89.5% of them did not find the prompt difficult to understand. In addition, 81.6% of all respondents found the key points suggested for the content of their writing useful. However, more respondents in Task 2 found it annoying to have the audience specified in the prompt (43%) than in Task 1 (31.3%). It seems that the students not only felt less comfortable writing a formal essay to their teacher, but that they were also more likely to be misled into writing a letter rather than an essay for their audience. Looking at the phrase in the prompt "write a formal essay… to your teacher" they

[39] This may be not because the tasks were difficult to do, but because to write in English *per se* was difficult.

seemed to focus on "to your teacher" and overlook the point "a formal essay", which specified the genre of writing. In this respect, specifying the audience in the formal essay writing worked negatively, which needed to be revised for the main study. Nonetheless, many respondents (69%) answered that they were able to demonstrate their writing ability properly through Task 2.

With regard to both Tasks 1 and 2, more than half of the respondents (58.7%) gave negative answers regarding the usefulness of specifying the minimum number of words (i.e., two hundred words). For 75.6% of them, on the other hand, the lack of choice of tasks did not matter or was felt to have a positive effect on fair rating. Finally, with regard to the question as to what kind of feedback they want on their writing,[40] 16.0% of them answered that they wanted individual ratings for writing features, and 61.4% wanted a teacher's comment on their writing as well as the ratings for this type of profile.

From the findings above, it could be said that both tasks were appropriate for the students. They found the wording of the prompts easy to understand, and although they had difficulty in deciding the content of their writing, the provided key points on content proved to be helpful, and they agreed that they were able to demonstrate their writing ability. Nevertheless, some parts of the prompts needed to be revised for the main study, namely, specifying the audience and the minimum number of words required. For the former, it seemed better not to specify a particular audience for the formal essay task for the reason mentioned above. As for the latter, it seemed better not to specify the minimum length as students reported that their focus was more likely to be on quantity than on quality. With revisions on these two points, I devised the writing tasks, Tasks 3, 4, 5 and 6 for the main study.

### 5.5.2 Writing tasks for the main study and the obtained scripts

The pilot study helped me to establish and devise the tasks (Tasks 3, 4, 5 and 6) for the main study. For the main study I changed the topics in order to exclude any practice effects. The new topics were also selected from textbooks in the same manner as for the pilot study.

---

[40] Although this question was not related to the investigation of the tasks, I included it as a quick way of finding their preferred method of getting feedback on their work.

As a result, I devised four tasks as follows:

- Task 3: write a letter to your foreign friend, informing him/her of places to visit in Korea;
- Task 4: write a formal essay explaining the advantages and disadvantages of using the Internet;
- Task 5: write a letter to a foreign friend describing your hobbies and talents;
- Task 6: write a formal essay explaining the advantages of learning English.

The same type of rhetorical specification was created for these tasks, apart from the fact that different cues were given for the content of each topic and no specifications were made regarding the audience (for formal essay genre only) and minimum length, both of which were included for Tasks 1 and 2.

As for the use of these tasks, Tasks 3 and 4 were essentially the main tasks, and were employed for the following three purposes: to obtain writing samples to be analysed as data for the development of the RS1; to be used as scripts to be assessed in order to investigate subjective holistic scoring and scoring using one of the existing rating scales (i.e., the FCE rating scale); and to be used as scripts to be assessed in order to investigate the practicality, reliability and validity of the Revised version of the new Rating Scale (RS2) (Groups A to G in Tables 5.1 and 5.2 below covered Tasks 3 and 4).

Tasks 5 and 6 were devised to obtain writing samples that the raters would be asked to assess to pilot the RS1. There were two reasons for employing different tasks (Tasks 5 & 6) from Tasks 3 & 4 in order to pilot the RS1. First, while comparison between the three rating schemes (subjective holistic scoring, scoring using FCE scale and the RS1) had to be made on the same writing data to avoid the effect of a text variable on rating, the piloting of the RS1 did not have to, since the results of the piloting would not be compared with those from the other rating schemes. Second, as the raters had to see the same writing data several times, I guessed that they might feel bored. I wanted them to find the assessment less tedious, at least for the piloting stage.

I obtained six hundred and sixteen scripts from the implementation of the main

tasks (Tasks 3 and 4). However, due to the limited availability of raters (see section 5.6.1 for detailed discussion), only three hundred and ninety scripts were actually used for this study. These were divided into Groups A, B and C. I also obtained another over six hundred scripts on Tasks 5 and 6, but because of the raters' workload for this study, I used only twenty of them[41], which were chosen by the researcher and put into Group H (see section 5.6.3 for further discussion). I allocated the obtained scripts to the three raters for this study, following their suggestions regarding how many scripts they would be able to assess. The scripts were grouped and allocated to the raters as follows:

Table 5.1 Grouping of the obtained scripts and allocation to the raters

| Scripts | | Grouping of the scripts | Rater |
|---|---|---|---|
| Division of scripts | Quantity of scripts | | |
| 3-A0201A~3-A0241A[42] | 41 | Group A | Rater A |
| 4-A0201A~4-A0241A | 41 | | |
| 3-A0401A~3-A0440A | 40 | | |
| 4-A0401A~4-A0440A | 40 | | |
| Subtotal | 162 | | |
| 3-A0101B~3-A0142B | 42 | Group B | Rater B |
| 4-A0101B~4-A0142B | 42 | | |
| 3-B0101B~3-B0140B | 40 | | |
| 4-B0101B~4-B0140B | 40 | | |
| Subtotal | 164 | | |
| 3-B1101C~3-B1132C | 32 | Group C | Rater C |
| 4-B1101C~4-B1132C | 32 | | |
| Subtotal | 64 | | |
| Total | 390 | | |
| Scripts for each band chosen by the researcher from all the samples on Task 5 | 10 | Group H | All the raters |
| Scripts for each band chosen by the researcher from all the samples on Task 6 | 10 | | |

As shown in Table 5.1 above, the obtained scripts for the main tasks, Tasks 3 and 4 were divided into three groups, Groups A, B and C, and some of the scripts were selected to be grouped into Groups D, E, F and G for the investigation of validity of

[41] Even if I had guessed that only some of the scripts would be used for this study when I was designing the research, I still would have had to ask all the subjects to do Tasks 5 and 6, as it would have been unrealistic to try to select some students to do these extra tasks.

[42] The ID number for each script was established for rating purposes. The first digit of the ID number (3 or 4) indicates the tasks, for example 3 for Task 3 and 4 for Task 4. The letter and numbers after the number 3 or 4 (e.g. A0415) indicate the grade, class and ID number of the writer being assessed. The final letter (A, B or C) indicates the rater. A, B and C mean Rater A, Rater B and Rater C respectively.

the RS2, as shown in Table 5.2 below.

Table 5.2 Re-grouping some of the obtained scripts for Tasks 3 and 4

| Scripts | | Grouping of the scripts | Rater |
|---|---|---|---|
| Division of scripts | Quantity of scripts | | |
| Ten scripts chosen from 3-B0101B~3-B0140B<br>Ten scripts chosen from 4-B0101B~4-B0140B | 20 | Group D | All the raters |
| Ten scripts chosen from 3-A0401A~3-A0440A<br>Ten scripts chosen from 4-A0401A~3-A0440A | 20 | Group E | Rater A |
| Ten scripts chosen from 3-B0101B~3-B0140B<br>Ten scripts chosen from 4-B0101B~4-B0140B | 20 | Group F | Rater B |
| Ten scripts chosen from 3-B1101C~3-B1132C<br>Ten scripts chosen from 4-B1101C~4-B1132C | 20 | Group G | Rater C |

As can be seen in Table 5.2, the scripts in Group D were selected from Group B, and those in Groups E, F and G were selected from Groups A, B and C respectively. Group D consisted of ten scripts on Task 3 and ten scripts on Task 4. These two sets of ten scripts were written by different students; that is, the writers of the scripts for Task 3 were not the same as those for Task 4. The use of Groups A to H will be discussed in section 5.6, where the research design is explained.

## 5.6 Design and administration

The main study following the Questionnaire I study and piloting writing tasks consisted of the following four phases:

Table 5.3 Summary of the study procedure

| Phase | Procedure | Data or materials used | | | |
|---|---|---|---|---|---|
| | | By the researcher | By Rater A | By Rater B | By Rater C |
| **Phase One**<br><br>obtaining data and investigati-ng available scoring schemes | (1)Obtaining scripts from all the subjects, and typing them up | All groups of scripts | N/A | | |
| | (2)Rating them, first according to subjective holistic scoring method, and at the same time keeping diaries on their assessment of each script for the purpose of a qualitative study | N/A | Group A | Group B | Group C |

| | | | | | |
|---|---|---|---|---|---|
| | (3)Assessing some of the script by subjective holistic scoring for the purpose of a quantitative study, i.e., intra-rater reliability of the scoring scheme | N/A | Group E (vs. scoring on the same scripts within Group A in (2)) | Group F (vs. scoring on the same scripts within Group B in (2)) | Group G (vs. scoring on the same scripts within Group C in (2)) |
| | (4)Assessing some of the script by subjective holistic scoring for the purpose of a quantitative study, i.e., inter-rater reliability of the scoring scheme | N/A | Group D | | |
| | (5)Doing a think-aloud for some of the scripts for the purpose of the qualitative study on the scoring scheme | N/A | Six specific scripts within Group D | | |
| | (6)Assessing using the FCE scale, and at the same time keepig diaries on their assessment of each script for the purpose a qualitative study | N/A | Group A | Group B | Group C |
| | (7)Assessing some of the script by subjective holistic scoring for the purpose of a quantitative study, i.e., intra-rater reliability of the scoring scheme | N/A | Group E (vs. scoring on the same scripts within Group A in (6)) | Group F (vs. scoring on the same scripts within Group B in (6)) | Group G (vs. scoring on the same scripts within Group C in (6)) |
| | (8)Assessing some of the script by the FCE scale for the purpose of a quantitative study, i.e., inter-rater reliability of the scoring scheme | N/A | Group D | | |
| | (9)Doing a think-aloud for some of the scripts for the purpose of the qualitative study on the scoring scheme | N/A | The six specific scripts within Group D | | |
| **Phase Two** developing the RS1 | (10)Analysing the scripts according to a coding scheme and developing the RS1 on the basis of the analysis | Groups A, B and C | N/A | | |
| **Phase Three** | (11)Using the RS1, keeping diaries | N/A | Group H | | |
| Piloting and revising the RS1 | (12)Answering a questionnaire to find out their opinions of the RS1 | N/A | Questionnaire III | | |
| | (13) Revising the RS1 into the RS2 on the basis of the feedback from the raters | Feedback obtained in (11) & (12) | N/A | | |
| **Phase Four** | (14)Assessig scripts using the RS2, and also keeping diaries for their assessment of each script for the purpose of a qualitative study | N/A | Group A | Group B | Group C |
| validating the RS2 | (15)Assessing some of the script using the RS2 for the purpose of a quantitative study in (19) | N/A | Group E | Group F | Group G |
| | (16)Assessing some of the script using the RS2 for the purpose of a quantitative study in (19) | N/A | Group D | | |

| | | | | |
|---|---|---|---|---|
| (17)Doing a think-aloud for some of the scripts for the purpose of a qualitative study of the RS2 (19) | N/A | The six specific scripts within Group D | | |
| (18)Investigating the practicality of the RS2 through a questionnaire study for the raters | N/A | Questionnaire IV | | |
| (19)Investigating validity of the RS2 | | | | |
| (19-1)For correlational evidence | N/A | Ratings on Group E obtained in (15) vs. ratings on the same scripts in Group A obtained in (14) | Ratings on Group F obtained in (15) vs. ratings on the same scripts in Group B obtained in (14) | Ratings on Group G obtained in (15) vs. ratings on the same scripts in Group C obtained in (14) |
| (19-2)For "G-study" (ANOVA) | N/A | Ratings on Group D obtained in (15) | | |
| (19-3)For MTMM | N/A | Ratings on Group D obtained in (15) vs. ratings on Group D obtained in (8) | | |
| (19-4)For questionnaire study on the raters | N/A | Questionnaire V | | |
| (19-5)For diary study | N/A | Diaries on Group A obtained in (14) | Diaries on Group B obtained in (14) | Diaries on Group C obtained in (14) |
| (19-6)For think-aloud study | N/A | Verbal protocols on the six specific scripts within Group D obtained in (17) | | |
| (19-7)For questionnaire study on students | N/A | Questionnaire VI | | |

Each phase will be discussed in detail in the following four sections.

### 5.6.1 Phase One

After a pilot study where the subjects were asked to complete Tasks 1 and 2 whilst the raters were asked to do assessment of the scripts on Tasks 1 and 2, keep a diary and do a think-aloud, the first task in the first phase was to obtain scripts from the subjects. For this, the subjects were asked to do two kinds of writing task, Tasks 3 and 4, over two sessions between September and October 2003. After doing these two tasks, they were asked to do Tasks 5 and 6, over two sessions between November and December 2003.

I asked T1, T2 and T3 to conduct these tasks in their classes (T1 was also Rater A, but T2 and T3 were not raters). Originally, it was intended that the students would have no prior information regarding the tasks, but the teachers advised me that this would be unfeasible because the limitations of the timetable would not allow all eight classes to complete the tasks at the same time. It would have been impossible to ensure that the students had the same conditions prior to completing the tasks, as

some would have completed them ahead of the others, which would result in unfairness. T1 was especially concerned about this; he decided to include the students' performance on these tasks in their final grade of the course, which he thought would further motivate them. Therefore, it was desirable that all the students had the same chances before completing the task in order to ensure that the grade given was based on the same performance conditions. Consequently, the teachers suggested that for fair management, the students should be informed of the topics in all the classes.

I fully understood the reasoning of the teachers, but remained concerned that the students might consult books or private teachers who would help them prepare for the tasks, which would have an influence on the features of their writing and in turn on the RS1. The teachers assured me that the students had prepared by themselves under similar circumstances, and that it would not impact on the demonstration of their abilities. Eventually, I agreed that the tasks would be announced to all the students before the writing sessions. The teachers informed the students of the topics a few days before each writing session. When each writing session started, a piece of paper with the prompt was handed out to every student, who then had fifty minutes to complete the task allocated. They were instructed to write by hand, rather than by using computers. Due to the time limitations of the curriculum there was a one-month gap between each task.

Once Tasks 3 and 4 were completed, the students' scripts were collected. The total number of scripts from these two tasks was six hundred and sixteen. Afterwards, I typed up the data in order to prevent bias due to the legibility of writing (Klein & Taub, 2005), since very neat writing gives a good impression to raters and so tends to get higher grades, whilst scribbled writing tends to get a lower grade than the content and English might deserve. In order to prevent this undesirable effect, I typed up all the scripts.

I then handed over the three hundred and ninety typed scripts to the three raters and asked them to rate these according to their own subjective judgement, without any formal rating scale. As mentioned in section 5.4, I had difficulty in finding raters who were willing to rate the scritps for this research. I could not find any more raters for the rest of the samples (i.e., two hundred and twenty-six written data) and the

three raters told me that they could not manage too many scripts due to the very heavy workload at school. As a result, I was obliged to give up the two hundred and twenty-six samples. Following the raters' opinion on how many scripts they would be able to assess, I allocated one hundred and sixty-two pieces (Group A) to Rater A, one hundred and sixty-four pieces (Group B) to Rater B and sixty-four pieces (Group C) to Rater C for rating. They were asked to rate on 1-to-6 bands, with 1 being the lowest (see section 6.2 for more details and results).

During the assessment, the raters were asked to make diary entries (Bailey, 1990; Bailey & Ochsner, 1983; Parkinson *et al.*, 2003) in order to reveal how they assessed the scripts and what they experienced during the assessment process. Although the diary-keeping was chosen as a way of tracking their thought processes while rating, it is recognised that this method has limitations in achieving this end. Parkinson *et al.* (2003) discuss this issue under the topic of the "'truth' of the diaries" as follows:

> Were the diarists writing what they really felt, or were they 'giving us what we wanted' [...]? Ultimate truth is of course never demonstrable (though its opposite is sometimes apparent), and the best one can hope for is 'a' truth rather than 'the' truth, representing the informants' beliefs, as formulated for others, but never all their beliefs [...] (p. 61).

Therefore, I decided to use this method, assuming that the diaries would be able to reveal, at least, "a" truth about the raters' introspection that would not be made available by any quantitative methods. The raters wrote these diary entries in Korean. However, as they had never made such diary entries before, I gave them instructions on diary-keeping and showed them the most informative models of those, which were written by Raters A and B during the pilot study in July 2003 (on Tasks 1 and 2). The models were to help them understand better what diary entries might look like, what they should be about, how they should flow, what the focus and aim of the diary entries were, and so on. Two of the model diary entries that they were shown are as follows:

> Band: 5
> I hesitated which band would be appropriate for this script between Bands 4 and 5. There are not grammatical errors in this script except one error in the use

of passive voice. As for the organisation, it could be understood well and clearly. The writer also used appropriate length of sentences and vocabulary. Unfortunately, any outstanding expressions which would point towards excellence were missing. So I am afraid it does not deserve Band 5. This notwithstanding, as there are no points which make it deserve a lower band than Band 5, I decided to assign Band 5 to it.

Band: 5
I hesitated between Bands 4 and 5 for this script. It is not well organised, with too many short paragraphs, and none of the paragraphs are developed enough. So this script gives me a sense of distraction. Additionally, it seems that there are many grammatical errors and inappropriate expressions in it. So I am afraid it deserves Band 4. However, given that it is very long and conforms to the genre format of formal essay, I decided on Band 5. Considering my rating process of this kind, it looks like I am influenced very much by length in rating. For me, I am likely to assign a better grade than the one a script really deserves if it is very long.

I handed over the typed-up scripts to the raters between mid-December 2003 and early January 2004. Rater A finished the assessment and diary-keeping according his own subjective judgement by July 2004, and Rater B by January 2004, but Rater C entered the research late, after several other raters had withdrawn, so he was handed the scripts in July 2004 and completed the assessment and diary-keeping in August 2004.

With an interval of at least three weeks after each of the raters finished the assessment and diary-keeping, I asked them to rate both Group D and one of Groups E, F and G. This time they did not have to make a diary. My purpose was to investigate both validity and inter-and intra-rater reliability (of course reliability belongs to the realm of validity, as discussed in section 3.3.1) of subjective holistic scoring, and to investigate the rating process through think-aloud procedure (see section 6.2 for more details and results). During the assessment of six specific pieces from Group D, the raters were asked to do think-aloud as well. This was to obtain more detailed and concurrent data on the rating process than could be given by the diaries. Since they were written retrospectively, in my opinion, the diaries could have captured simple/short comments on the rating process, and to obtain fuller remarks and comments, to employ a think-aloud method, which would be done concurrently with the assessment, seemed helpful. According to previous studies, the think-aloud method is well suited to research on the rating process (e.g., DeRemer, 1998;

Vaughan, 1991), allowing raters to pour out their thoughts during the assessment, hopefully with less 'reduction'. My aim was to use both of these methods (think-aloud and diary) to obtain rich data on the rating process. However, it must be admitted that neither kind of data can give a full description of the rating process, because even think-aloud cannot give a full picture of mental processes and that, as Glendinning and Howard (2001) claim, the think-aloud method is "artificial" in that participants are required to voice their thoughts (p.42).

Having finished the rating, diary-keeping and think-aloud according their own subjective judgement, the raters were invited to use the FCE scale to assess the same scripts. It was pointed out in Chapter One that published rating scales were not satisfactory for the given context, and that it was my intention to investigate this empirically. I asked the raters to use one of the published rating scales, the FCE scale, and assigned Group A to Rater A, Group B to Rater B, and Group C to Rater C. As with the subjective holistic scoring, they were asked to use the rating scale and keep a diary on their rating process. Rater A finished this in September 2004, Rater B in March 2004 and Rater C in September 2004. Again, as with the subjective holistic scoring, they were then invited to assess Group D for the investigation of inter-rater reliability. Groups E, F and G were assigned to Raters A, B and C respectively for both the investigation of intra-rater reliability and a think-aloud using the scale (see section 6.3 for more details and results).

### 5.6.2 Phase Two

The scripts obtained from Phase One were analysed to develop a rating scale for this study. For this, a coding scheme was developed between July and early December 2004. The development was based on literature review, the definition of writing ability implicit in the course objectives and the preliminary analysis of the obtained scripts (see section 7.3.1 for more details).

After the scheme was devised, the next step before it could be used for coding of the scripts was quality control of coding data with the scheme. This was done in December 2004. According to Fulcher (2003), there is no one-to-one mapping between data and the explanation of errors and characteristics in the data. Therefore, "data-driven scale development needs to employ methods that control the quality of

data coding and the interpretation" (p.102). Of the two methods he suggested for this in his book, namely, double (or triple) blind coding and discriminant analysis, I chose the former. Working separately, a Korean postgraduate student in applied linguistics and I used the coding scheme for blind coding of twenty randomly chosen of the obtained scripts and then coded them before measuring the degree of agreement between us, i.e., the inter-coder reliability. To measure agreement, I calculated the coincident codings between us divided by all coding occasions per script multiplied by 100 (%), and then counted the mean of agreement percentages for all twenty scripts. In this investigation, the agreement was 88.1%. Therefore, I felt justified in using this coding scheme (see section 7.3.2 on the coding scheme).

After establishing the coding scheme, I coded three hundred and ninety scripts between late January and late February 2005. I then used Crosstabulation, Test of independence, i.e., chi-square and Frequency analysis to identify features discriminating between neighbouring bands. On the basis of the features thus identified, I developed the RS1 in April 2005 (see section 7.5 to 7.6.1 for more details of the rating scale development procedure).

### 5.6.3 Phase Three

The RS1 was piloted in the third phase between late April and May 2005. This was an opportunity to try out and revise the scale. For this phase, Group H, i.e., a part of the scripts on Tasks 5 and 6 were assessed. The three raters were not given all of the written data from the two tasks, but ten scripts for Task 5 and ten scripts for Task 6 that included cases of all the bands that I had judged according to my own subjective judgement. Ideally, it would have been good to have as much data as possible for this phase, but there was a practical limitation that the raters felt burdened with their extensive workload. I thought the quality rather than the quantity of feedback as more valuable at this stage, given that it was only a trial stage and that its aim was to obtain feedback in order to revise the scale. Accordingly, I had a small number of scripts for this stage, but asked the raters to give very detailed feedback, keeping a detailed diary and to respond to a questionnaire, composed of mainly open-ended questions, regarding the scale (i.e., Questionnaire Ⅲ in Appendix 10). Taking account of the diary entries and the responses to Questionnaire Ⅲ, I revised RS1 to

produce RS2. I tried to improve the wording so as not to be repetitive or contradictory, and changed the format/layout of each cell in the scale. I made both general features and features distinguishing from lower or higher neighbouring bands visually clear, with general features on the top of each cell and distinctive features indicated by arrows such as '▲' and '▼' before each descriptor. This was to help the raters grasp the general characteristics of each band first and decide between possible candidate bands for a script. In addition to these revisions to the scale, I tried to improve the raters' manual to make it more user-friendly by including detailed explanations and concrete examples for each descriptor, and suggesting how to calculate the total band from three scores on three assessment categories, according to one rater's suggestion.

### 5.6.4 Phase Four

During this phase, the RS2 was applied to a practical assessment. Between July and August 2005, Raters A, B and C used this rating scale to assess Groups A, B and C respectively. As in Phase One, they were also asked to keep a diary during the assessment.

After they finished assessing Groups A, B and C, they were asked to answer a questionnaire, Questionnaire IV (see Appendix 11) that I had devised to investigate the practicality of the scale.

About two weeks after completing the assessment, each rater was invited to rate Group D and one of Groups E, F and G, and to do a think-aloud using the RS2. In addition, two questionnaire studies were carried out, one for eighty students (see section 8.4.8 for more details on these students) and the other for the three raters. All the data obtained in this phase (the ratings of various script groups, verbal protocols, diaries and questionnaire answers) were employed to investigate the validity of the RS2 through seven methods: correlational evidence, the MTMM, a "G-study" (ANOVA), two questionnaire studies, a diary study and a think-aloud study. After the validation procedure, the scale was compared to the FCE scale to find the differences between them.

## 5.7 Summary

This chapter has dealt with how the present research was designed and conducted. The study was on Korean high school students in the first and second year at Kwacheon foreign language high school in Korea. They were asked to do four kinds of writing tasks in all. Two of them were to obtain writing samples that would be used to develop a rating scale, then employed as the subject of an investigation of subjective holistic scoring and the FCE scoring, and finally to investigate the practicality, reliability and validity of the RS2. The other two tasks were to obtain writing data with which I would pilot the RS1. These tasks were completed in class over four sessions, and the students were informed of the topics in advance of each session.

After the samples were obtained, I typed them up and they were assessed by three raters who worked at various high schools in Korea. They were asked to assess the data on Tasks 3 and 4 using three rating schemes: subjective holistic scoring, the FCE scale and the RS2. Before the stage of using the RS2, the raters used the RS1 to assess twenty scripts generated from Tasks 5 and 6, and were asked to give feedback on it. I was able to revise the RS1 on the basis of their responses. During these assessment periods, the raters were asked to keep a diary that would illuminate their rating process and show both the characteristics and problems of the three rating schemes.

To develop a coding scheme that were needed for the development of the RS1, the obtained scripts were used, literature was reviewed and the definition of writing ability implicit in the English Writing course was considered. After going through the stage of quality control in coding data using the scheme, I coded three hundred and ninety scripts according to the scheme. All of these were then analysed statistically to find the features discriminating between neighbouring bands, which were the basis of the development of the RS1.

After the RS1 was developed, piloted and revised, it was investigated for its quality, namely practicality, reliability and validity.

The following chapters will give a more detailed description of the procedures and results of the research procedure from Phases One to Four described in this chapter. Chapter Six will deal with the detailed procedure and results of applying

existing rating schemes (i.e., subjective holistic scoring and the FCE scale), Chapter Seven will cover Phases Two and Three, and Chapter Eight will cover Phase Four.

# CHAPTER SIX. INVESTIGATION OF EXISTING RATING SCHEMES: SUBJECTIVE HOLISTIC SCORING AND THE FCE RATING SCALE

## 6.1 Introduction

As stated in Chapters One and Five, I employed three empirical methods to explore the need to develop a rating scale for the English Writing course at FLHSK. First, I asked English teachers in FLHSK how they assessed their students' writing and whether they thought a rating scale needed to be developed for this (see section 5.2 for more details). Second, I empirically examined subjective holistic scoring, and third, I examined one of the published rating scales, the FCE scale. Thus, the three raters for this study were first asked to rate the obtained scripts according to subjective holistic judgement and then rate using the FCE rating scale. To find out what they experienced during the assessment, I asked them to keep a diary and do a think-aloud separately. This chapter, therefore, will evaluate the two rating schemes through both quantitative analysis of the ratings, and qualitative analysis of the rating process through the diary study and the think-aloud study.

## 6.2 Subjective holistic scoring

### 6.2.1 Scoring procedure

The procedure for subjective holistic scoring was for the three raters to rate the scripts (Group A for Rater A, Group B for Rater B and Group C for Rater C) on a band of 1 to 6, with Band 1 being the lowest level. Their ratings at this stage were to be used for two purposes: (1) to allow all the scripts to be sorted into six groups and then analysed using the coding scheme to find the distinctive features of each band; (2) to obtain first ratings, which would subsequently be compared with their second ratings for the investigation of their intra-rater reliability (see Table 5.3 for details). They were also asked to keep a diary of the assessment of each script in Korean.[43]

Apart from having to assess on a 6-point-scale, each rater had their own rating standard. This allowed them the freedom to reflect their experience and preferences

[43] Whilst Rater C was asked to make diaries for all the scripts (i.e., Group C), Raters A and B were only asked to keep a diary for half of Groups A and B respectively as each of them was more than twice the size of Group C, in order to reduce their diary-keeping workload. This means that there were fewer diaries than ratings.

as teachers and their philosophy of writing. Thus, the raters were asked to decide which aspects of the scripts to rate – content, vocabulary, organisation and so on, and what level of proficiency each band would require.

With regard to diary-keeping, I asked the raters to consult the most informative diary entries made by Raters A and B during the pilot study stage (see section 5.6.1 for the examples of model diary entries).

About three weeks after rating their own scripts (that is, Groups A, B and C respectively) according to their subjective judgement, I asked all the raters to assess Group D and do a think-aloud for the six scripts in Group D. Rater A was also asked to assess Group E, Rater B to assess Group F and Rater C to assess Group G. I also referred them to the most informative models of think-aloud among those recorded by Rater B and a rater who later withdrew but functioned as Rater C in 2003 during the pilot study stage. I distributed not only cassette tapes containing these examples, but also guidelines on how to do think-aloud, which I drew up on the basis of Ericsson and Simon (1993) and Green (1998). After referring to the examples and guidelines, they recorded think-alouds by themselves in Korean without the researcher being present.

As a result, I obtained three hundred and ninety ratings, two hundred and twenty-four diary entries and eighteen recorded think-aloud protocols when the raters finished assessing the scripts according to their subjective holistic judgement. I transcribed the think-aloud protocols, and then analysed the ratings quantitatively and the diary and think-aloud data qualitatively. I will discuss the results of the analyses in the following sections.

### 6.2.2 Quantitative analysis

The quantitative analysis looked at two aspects of reliability: intra-rater reliability and inter-rater reliability. This analysis was carried out because inter- and intra reliability figure among the highest priorities in the quantitative study of ratings. The investigation was conducted from the viewpoint that reliability is one element of evidence of validity, rather than seeing reliability and validity as separate concepts, of which reliability is the most important, as discussed in sections 3.3.1.1 and 3.3.1.2.

These studies used the ratings from Group D for inter-rater reliability and those

from Groups E, F and G for intra-rater reliability (which were compared with the ratings of the same scripts in Groups A, B and C respectively). They were analysed with SPSS software. The results for intra-rater reliability are presented in Table 6.1 below.

Table 6.1 Intra-rater reliability for subjective holistic scoring

| | Rater A1 * Rater A2 | Rater B1 * Rater B2 | Rater C1 * Rater C2 |
|---|---|---|---|
| Intra-rater reliability | .262 | .761** | .780** |
| N | 20 | 20 | 20 |

NB. In this table, Rater A1 and Rater A2, for example, mean Rater A's first rating and second rating respectively.

As can be seen, Raters B and C showed high intra-rater reliability, significant at the .01 level. This was not the case with Rater A, which implies that while Rater A's rating was not consistent, Raters B and C had relatively firm subjective criteria which they applied fairly consistently.

Inter-rater reliability was as follows:

Table 6.2 Inter-rater reliability for subjective holistic scoring

| | Rater A | Rater B | Rater C |
|---|---|---|---|
| Rater A | 1.000 | .686** | .703** |
| Rater B | .686** | 1.000 | .613** |
| Rater C | .703** | .613** | 1.000 |

Table 6.2 above shows that inter-rater reliability was relatively high and significant at the .01 level. Notably, the inter-rater reliability involving Rater A, whose intra-rater reliability is not significant, is also quite high. Given that the intra-rater reliability coefficient is the upper limit of estimated inter-rater reliability, these high inter-rater reliability coefficients is beyond expectation.

In summary, the inter-and intra-rater reliability for subjective holistic scoring appeared to be quite high, apart from Rater A's intra-rater reliability, which means that subjective holistic scoring could be suitable in terms of quantitative analysis and could have content-related evidence of validity. However, this finding needed to be examined through qualitative methods. In the next section, therefore, I will discuss the raters' rating process.

### 6.2.3 Qualitative analysis: diary and think-aloud analysis

#### 6.2.3.1 Data analysis procedure

For the qualitative analysis of the diary entries and think-aloud protocols, I followed a strategy which "involve[d] scanning the data for categories of phenomena" (Goetz & LeCompte, 1984: 80) and examining the data "for meaningful themes, issues, or variables to discover how these are patterned, and to attempt to explain the patterns" (Johnson, 1992: 90). Thus, I read the obtained diary entries and verbal protocol data repeatedly both during and after transcription, noting and coding both recurrent themes and comments salient to the purpose of this research.

The analysis revealed two main themes and their sub-features as salient to the purpose of these diary and think-aloud studies: observed patterns and tendencies, and problems with subjective holistic scoring.[44] They will be discussed in detail in the following sections. They could be discussed according to each rater in which manner Parkinson *et al.* (2003) did, but as there are many aspects that are common to all raters for this study, I will discuss them as a group here in order to avoid too much repetition, specifying any points that apply to an individual rather than the group.

#### 6.2.3.2 Observed patterns and tendencies in subjective holistic scoring

First, all three raters mentioned content and grammar. As for their attention to content, the same phenomenon was found in Brown (1991), Freedman (1981) and Vaughan (1991). Their focus on grammar o these two features concurs with McNamara's (1990) view, quoted by Storch (1993), that raters' attention to grammatical accuracy is "very deep-seated" (p.23). In Storch's (1993) study, it was revealed that grammatical accuracy and appropriacy were found to be one of the two most influential categories in rating.[45] The following diaries demonstrate the same

---

[44] The analysis of the diary and think-aloud protocols here was theme-based rather than statistics-based. The latter was not seen as desirable for this study, given the raters' diary-keeping behaviour. Reviewing the raters' diary entries, it was obvious that as they rated, they omitted some comments that were repeated in the scripts since they felt bothered and bored by writing repeated phrases/sentences. Thus, as they went on making diaries, the entries became shorter and shorter. Therefore, it was not seen as appropriate to do statistical analysis of the comments and themes that emerged, so I did theme-based analysis of them here. Since there were only six think-aloud protocols for each rater and the same reasoning seemed to apply as for the diary analysis, it was not deemed desirable to do statistical analysis of the comments in the protocols, either.

[45] The other was lexical accuracy and appropriateness. Therefore she argues that these two aspects of grammatical and lexical accuracy and appropriacy need to be developed in classes (Storch, 1993).

tendencies:

ID: 3-A0415A[46]

Band 3

This script could be good because it is fairly long and well organised. But as I looked at it more closely, I found a major error in it. It is that this script deals with the myth of Ulsan Rock rather than the required content from the prompt--places in Korea which are worth visiting. Additionally, it has basic grammatical errors and does not include connectors which can be seen in other students' writings. Therefore, I think Band 3 is most appropriate for this script.[47]

ID: 3-B0118B

Band 4

This writer seems to have made an effort, as the script is very long. The writer tried to use a wide range of vocabulary and content, which creates a good impression. However, the writer kept making errors in the use of relative pronouns and the spelling of very basic words. Taking into consideration the very frequent errors in these aspects, it deserves Band 3. However, due to the good aspects mentioned above, I generously assigned Band 4.

ID: 3-B1119 C

Band 4

This script introduces Jeju island, Seolak Mountain and Kyungbok Palace in Korea. It communicates fairly well. However, the use of slang such as *cos* (this should be corrected into *'cos)* and *gonna* does not look good, given not only that they are being used in written language, but also that the writer is a student. But the content is good so I assigned Band 4, despite more than ten errors in this script.

Looking at both the diaries above and their other diary entries, it can be seen that while Rater C rated only content and grammar (and sometimes expression for advanced levels), Rater A paid attention to grammar, content, length, paragraphing, genre format, organisation and complexity, and Rater B to intelligibility, length, paragraphing, genre format and organisation, in addition to content and grammar. They chose their own assessment categories depending on the level of a script, but in

---

[46] All the information for the diary entry is given here, including the ID number established for rating purposes, the rating result and the comment of the rater. In the ID number, the first digit (3 or 4) which of the two tasks for this study is involved: 3 is for Task 3 (to write an informal letter to a foreign friend describing the places in Korea which are worth visiting), whilst 4 is for Task 4 (to write a formal essay explaining the advantages and disadvantages of Internet use). The letter and numbers after the number 3 or 4 (e.g. A0415) indicate the grade, class and ID number of the writer of the script being assessed. The final letter (A, B or C) indicates the rater. A, B and C mean Rater A, Rater B and Rater C respectively. All the comments in a diary entry are shown in this example, but where necessary I have omitted sections that are irrelevant to the topic under discussion and added emphasis in bold.

[47] All the diary entries and think-aloud protocols obtained for this study were originally made in Korean. I have translated them into English.

every case they considered grammar and content.[48]

Second, the raters demonstrated 'central tendency' (e.g., Brown, 1995; Lumley, 1995; Vaughan, 1991) in that they avoided assigning scripts to the highest and lowest bands.[49] In fact, only one of the three hundred and ninety scripts was assigned to Band 1. The raters avoided assigning not only Band 1, but also Band 2, as can be seen in the following diary entry and verbal protocol by Rater A:

> ID: 3-A0414A
> Band 3
> Although the writer seemed to try to write smoothly as his/her thoughts flowed, what s/he has actually produced is a script that looks illogical. What is worse, the writer does not attend to paragraphing and organisation at all. The length is also insufficient and it is filled with many grammatical errors. I marked it as Band 3, but to be honest, I would like to put it in the lower band, Band 2.

> ID: A-INT-TA-03[50]
> Band 3
> Um.... This script is less than half the minimum length. So, the maximum band would be 3 for this....
> Hi, I'm going to recommend you a great place to visit in Korea. If you search..... in the map, you'll find a big island in very south of Korea. This doesn't conform to letter format... there's no way this can get a good band ... It called Jeju island. Um... Passive voice here is wrong.... (*the middle part of the rater's comment is omitted*).. I hope to see you in Jeju 2 years later. I'll write you more about Jeju later.
> Bye.
> With love, _____.[51]
> The closing part is good because it ends with *with love*, but...it doesn't conform to letter format... It is too short...and not accurate... No signpost for organisation...Maximum band for this would be 3, I suppose... Ah.... Um... It would be too harsh to give Band 2 to this... Well.. I will just assign Band 3.

---

[48] As for the order in which they considered these aspects, Rater A considered length, paragraphing, genre format, organisation, then content and grammar, whilst Rater B did not show any specific pattern with regard to order.

[49] This tendency is also called 'shrinkage factor' (Storch, 1993).

[50] As with the diary entries, I include all the information for the think-aloud protocol, including ID number of the writing sample, the rating result and the rater's comment. In the ID number, the first letter (A, B or C) is for rater identification, and the following letters and numbers identify the writing samples for think-aloud. For the sake of clarity, I used three types of font in the text. Arial font is for the rater's comment, Times New Roman is for the script being assessed, and Italic indicates the rater's behaviour or researcher's comment. Where necessary, I omitted irrelevant sections and added emphasis in bold. I did not include phonetic information for the comment, but indicated pauses lasting several seconds using "....".

[51] To avoid raters' recognition of the writer of a script being assessed, names in opening and closing parts in the scripts on Task 3 were deleted, when it was typed up.

This central tendency does not seem to be limited to the raters in this study. Indeed, as McNamara (1996) points out, it is found very widely. According to him, raters differ systematically in the range of scores they normally use out of the whole range available in a scale. Some raters show central tendency, and avoid scores at each end of the scale, while others avoid the scores in the middle of the scale to try to see differences between candidates. The raters in my study appeared to follow the first of these two patterns. It seems that they tend to avoid the lowest band(s) as long as students write something, since Korean teachers generally take account of the effort involved in a piece of writing.

Third, the raters focused on different assessment categories depending on the level of the scripts. Raters A and B paid attention to various assessment categories during the assessment, depending on the level of the scripts they were assessing. They mainly focused on length and/or grammar when they were assessing low-level scripts such as Band 2 scripts, as can be seen in the following diary entries:

> ID: 4-B0104B
> Band 2
> There are grammatical errors in almost every clause. Most of them are caused by the use of inappropriate vocabulary, and consequently there are many clauses which are unintelligible. Additionally, the writer makes errors in the use of verbs. That is, since he/she does not know the exact meaning of the verbs he/she makes errors by either using a preposition wrongly or omitting it after a verb. What is worse, it is too short. I suppose there are few aspects that would help this put script in a good band.

> ID: 3-A0423A
> Band 2
> This script is extremely short. It's only one paragraph long, and what's worse is that it contains many grammatical errors. So, I have no choice but to assign Band 2 to this script. But if this kind of script deserves Band 2, what kind of script deserves Band 1?

However, when the raters were assessing Band 3- and Band 4-level scripts, the noticeable features to which they attended were organisation and intelligibility (in addition to content and grammar), as is demonstrated in the following diaries:

> ID: 3-B0127B
> Band 3
> This script is at the very middle level, I think. Whilst communication is relatively good on the whole, it has errors in terms of differentiating between singular vs. plural, tense and word classes. The errors, however, are local rather than global, so

they do not affect the communication of clauses. So I marked it as Band 3.

ID: 3-A0427A
Band 4
This writer does not care about formal aspects of writing such as paragraphing, so it should get a low level, compared to the previous script. But once I read it I found that the flow of discourse was very smooth. Also, although there are many grammatical errors, they are not serious. If it were well paragraphed, I would assign Band 5, but since it was not, I gave Band 4.

For Bands 5 and 6, it is noticeable that all the three raters mentioned expressions and sentence structure which looked natural and like native English. This was even the case with Rater C, who consistently assessed scripts mainly in terms of content and grammar, as demonstrated in the following diary entries and a verbal protocol:

ID: 4-B1123C
Band 6
In addition to the content, this script is absolutely excellent. The use of vocabulary and idioms is very fluent and at university level. Although there are errors in the use of idioms like "*take into account*" and "*make good use of*", they appear to be mistakes. Generally the writer has a good command of advanced expressions, and I am sure that his/her English will develop further.

ID: 3-B0122B
Band 5
This script develops very smoothly. Furthermore, it includes very appropriate expressions which are seldom used in other students' writing. The script communicates very well, in spite of not using long and complex sentences. So I assigned a high band, Band 5.

ID: B-INT-TA-03
Band 4
(*The rater starts reading the whole script*)
Hi, I'm going to recommend you a great place to visit in Korea. If you search in the map, you'll find a big island in very south of Korea. It called Jeju island....(*the middle part of the script omitted*)...And make sure to buy lots of tangerines because it is very delicious. I hope to see you in Jeju 2 years later. I'll write you more about Jeju later.
Bye.
With love, _____.
Generally it doesn't seem well organised, and.... due to unnecessary paragraphing, it doesn't seem to me that the content is good either... **There are no advanced expressions and vocabulary...** It looks as if this script is at a level where the communication of basic meaning is possible. So.. um... Band 4 appears to be suitable for this script.

The diary entries and protocols shown above reveal that the raters commonly require advanced and rather sophisticated as well as accurate English for the highest bands. They do not seem to consider accurate English as sufficient for the highest bands.

This finding, that raters pay attention to different assessment aspects depending on the level of English in the script, is similar to a finding in Pollitt and Murray's (1996) study. They investigated which performance features judges paid attention to during the assessment of speaking performance. They found that when the judges were assessing the performance of high-level candidates, they did not mention the features that they mentioned during the assessment of low-level candidates, and vice versa, which implies that they focused on different assessment aspects depending on the candidate's level of proficiency.

Finally, in many cases the raters hesitated between two bands before assigning a script to a particular band, and were unlikely to make a confident decision on the first reading. The following think-aloud protocol and diaries support this point:

ID: A-INT-TA-02
Band 5
To _____,
Hi, _____. How have you been?
....(*the middle part of the script and the rater's comment are omitted*)... See you in this winter. Take care.
Um.... first of all it's long enough .... Good paragraphing using words such as '*first*' and '*second*'.... Um... Conforms well to genre format of letter ... um... There are many ungrammatical sentences... Well.... This script could be assigned Band 4 or 5... It is not good enough for Band 5 or 6...I suppose this is probably Band 4..... But its format and content are excellent... Even though there are many grammatical problems, taking those good points into consideration, I will put it in Band 5.

ID: 3-A0403A
Band 4
I had trouble deciding between Bands 3 and 4. This script is fairly well organised, but not every paragraph is equally developed. What is worse is that it contains many grammatical errors, so it looks bad. Still, its detailed description of cultural assets is very laudable. In addition, this writing is also good in that it is organised although not perfectly. Taking these two good points in consideration, I decided on Band 4.

ID: 3-B0123B
Band 2
I hesitated between Bands 2 and 3 for this script. I thought it could deserve Band 3

in that it communicates relatively well. After I read this script once more, I found there were some colloquial contractions which were rarely used in written language and there were some Korean words which were spelled in English without any effort to find approximate and possible English words. Furthermore, it has very basic errors in spelling and punctuation (i.e., full stop). Therefore, I assigned the lower of the two possible bands Band 2.

As can be seen in the verbal protocol above, the raters were having trouble deciding between two candidate bands, which opened up the real possibility of inconsistent assessment. However, according to the quantitative analysis of the raters' ratings of subjective holistic scoring, except for Rater A's intra-rater reliability, they were found to be fairly high. Even though general (overall) reliability turned out to be high and may look fairly consistent, some individual test-takers may have been assessed unfairly due to this uncertainty.

In sum, all the raters commonly focused on content and grammar (accuracy) during the writing assessment, but paid attention to different writing features depending on the level of writing. For example, they focused on length and/or grammar for the lowest level, and on advanced expression for the highest band. In other words, they appear to expect writing of the highest level to contain advanced vocabulary and sentence structure. Additionally, they sometimes hesitated between two candidate bands before deciding one of them, and rarely used the lowest bands, 1 and 2. In the next section I will discuss the problems with subjective holistic scoring revealed in the diary entries and think-aloud protocols.

#### 6.2.3.3 The problems with subjective holistic scoring

I identified six problems with subjective holistic scoring. First, the raters sometimes assessed an aspect of scripts that did not correspond with any recognised measure or component of writing ability. This is evident in the following diary entries and a verbal protocol:

ID: 3-A0405A
Band 3
I hesitated between Bands 2 and 3 for this script. It could be Band 2 in that it is only half as long as the other scripts, is poorly organised and shows poor sentence construction and grammar. There are few well constructed sentences without

grammatical errors. **This notwithstanding, I assigned it Band 3, because it looks as though the writer has made an effort.**

ID: 3-B0118B
Band 4
The length of this script indicates that the writer made a real effort. The writer tried to use such a range of vocabulary and content that it gives me a good impression. Nevertheless, the writer kept making errors in the use of relative pronouns and in very basic spelling of words. Taking the very frequent errors in these aspects into account, it deserves Band 3. However, given the good aspects mentioned above, I generously assigned Band 4 to it.

ID: C-INT-TA-02
Band 4
To _____,
Hi, _____. How have you been?
…..(*the middle part of the script is omitted*)
See you I this winter. Take care.
…..(*the first part of his comment on this script is omitted*).. So.... this writer uses several examples of slang and doesn't seem to know how to use articles... And...there are many errors with *much* and *many*... many errors.... As for content, ... um.... **this script is written very long and very sincerely, so I wish to give a high band, but...** in spite of good organisation, there are too many grammatical errors.... So Band 5 would be appropriate for this.... Um.... For my criteria, less than 5 grammatical errors in a script could be assigned Band 5.... But this script has more than 5 errors... Even though the content is excellent, I am sorry, but I have to give Band 4 to this script....

As these diary entries and think-aloud protocol show, the rater took account of the effort made by the writer being assessed, i.e., whether he/she was trying to do his/her best, which is not recognised as a component of writing ability by language testing researchers. One reason for this phenomenon might lie in the fact that they were assessing according to their subjective impression, rather than using any formal rating scales that give explicit guidance on which features raters need to consider.

Second, the raters were influenced by the order in which the scripts were graded. Their rating was likely to be influenced by the level of the preceding script, as revealed in the following diary entries:

ID: 4-B0115B
Band 2
This script is less than the minimum length, consisting of only a few lines that look very insincere. Compared with the previous one, however, the sentences are intelligible and there are few grammatical errors.... (*the rest of this diary is omitted*).

ID: 3-A0409A
Band 4
This script is well organised and describes the touring places well. Additionally, there are fewer grammatical errors than in the previous one. ... (*the rest of this diary is omitted*).

This may also occur to some extent even when a formal rating scale is employed. However, I think that it would be less likely when a formal rating scale is used, because the raters have an invariant, visible, and substantial standard to consult with a formal rating scale. Therefore, it appears that there is less likelihood of being influenced by the order variable than in the case of subjective holistic scoring, where the raters have only an abstract standard in their mind.

Third, the raters were not always confident of their rating. After they assigned a band to a script, they sometimes wondered if their rating was appropriate for it. This is shown in the following diary entries:

ID: 4-B0107B
Band 4
If I assign Band 4 to this script, I think it would be very harsh. This is because this script is fairly good in terms of length, organisation and content. However, its weak point is that the writer tried to use a variety of expressions and this resulted in some awkwardness. While I read through the first half of the script I thought these awkward expressions would not negatively affect the grade, but in the second half, they negatively affected my impression of it and in turn the grade, because they sometimes hindered communication and intelligibility, and the former is one of the most important assessment categories for me. However, I wonder if it is fair to mark it down merely because of a few sentences that are not intelligible, even though the script communicates well on the whole. I am not sure whether the grade I assigned is fair.

ID: 3-A0419A
Band 4
I assigned Band 4 to this script. When I first read it, I thought it might not be good enough to get Band 4 because it was not paragraphed, so I had difficulties understanding the content. But when I had a closer look at it, I found the sentences in the script were accurate or grammatical and it described Kyungju in detail to such an extent that later I even thought it deserved Band 5. As a result, I am unsure if Band 4 is appropriate for this script. Nonetheless, I assigned Band 4.

ID: 4-B1128C
Band 4
It is not easy to grasp the content of this script even though its vocabulary, idioms and grammar are excellent. I find sentences like "*Internet contribute....culture. That's anyone gets....by one click*" really hard to understand. Because of these sentences which cannot be understood at a glance, I can't give the top level of

bands such as Bands 5 and 6. Even so, I'm not sure whether Band 4 is appropriate for this script because it seems a bit harsh.

As the diaries reveal, the raters were at times unsure of their rating. To some extent, this could result from the fact that they relied on their impressionistic judgement without using any formal rating scales. Consequently, an external confirming tool for their judgement might be helpful, especially when the level of the scripts is ambiguous. This could be partly (but not entirely) resolved by using a formal rating scale that explicitly indicates the proficiency level required for each band.

Fifth, it seems that the raters lacked clear criteria for some bands, and were not always sure of the proficiency level required for them by their own judgement scheme. The fact that Bands 1 and 6 were rarely used is probably partly due to their central tendency (see section 6.2.3.2 for more details), but may also be because they did not have clear criteria for them (particularly Band 1), as observed in the diary entries below.

ID: 3-A0423A
Band 2
This script is extremely short – just one paragraph long. And not only is it short, but it contains many grammatical errors, so I can't help but assign it to Band 2. **But if I mark it as Band 2, I'm not sure which scripts deserve to be put in Band 1**.

ID: 3-A0426A
Band 6
This script is excellent - the best so far. It is fantastic in organisation, genre format and paragraphing, and each point is equally developed and written very clearly. There are just a few grammatical errors. **But even though it is written well, I hesitated in assigning Band 6 to this script because I'm not sure whether it is really good enough to deserve Band 6, given that Band 6 is the highest band and would mean "perfect". Even though I am unsure about this, I assigned Band 6 to this script because it is the best bit of writing I have seen so far.**

Unlike subjective holistic scoring, formal rating scales explicitly indicate the level of each band. Hence, I believe that this difficulty could be resolved to some extent if a formal rating scale is used.

Finally, the ratings given by subjective holistic scoring, which were presented in a single number, did not reflect what the script really looked like or how the rater had arrived at the score. As is evident in the following diary entries, the rating process

was a kind of negotiation between various assessment categories in the raters' subjective rating scheme, which resulted in the script being assigned to a particular band. Thus, two scripts could end up the same band (such as Band 4) even though they were judged so for different reasons.

> ID: 3-A0427A
> Band 4
> This writer does not seem to care about formal aspects of writing such as paragraphing, so I thought it would deserve a lower level than the previous script. Once I read it, however, I found that it flowed very smoothly. In addition, although there are many grammatical errors, they are not serious. If it had been well paragraphed I would have assigned it to Band 5. But since it is not, I put it in Band 4.

> ID: 3-A0431A
> Band 4
> I hesitated between Bands 4 and 5 for this script. First of all, it is much longer than the minimum length. Also it is well organised, having good paragraphing and conforming to the genre format. Additionally, the grammatical errors are local and not serious. On the other hand, its content focuses excessively on the programme at Kimchi museum, which leads the content astray. Reflecting this weak point, I assigned it to Band 4.

These diary entries by Rater A show how two scripts were assigned to Band 4 for different reasons. The first one could have been assigned to Band 5, but was rated in Band 4 because the paragraphing was not good, even though it flowed well. We can see in the second entry that the rater had trouble deciding between Bands 4 and 5, and ultimately assigned the script to Band 4. In this case, however, it was content that prevented it from being put in Band 5. Thus, although the two scripts were assigned the same band, the process of negotiation that determined which band they were put in varied. This is consistent with the finding in Brown (1991), O'Loughlin (1992) and Shi (2001) (see section 4.6 for details). Hamp-Lyons (1991b) also mentions this point that "there are many ways to get a '7'" (p. 257). Without detailed accompanying comments, the ratings of subjective holistic scoring do not indicate the good and bad points of each script. Having been formulated through the unobservable process of subjective holistic scoring, such ratings are therefore not helpful for learners, who find more detailed information on their writing helpful for future improvement (the Korean students also requested detailed feedback, as discussed in section 5.5.1). If the scale in use is a holistic scale rather than an analytic scale (or multi-trait scale), this

problem is unavoidable even if a formal rating scale is adopted. In this regard, it is desirable to employ an analytic type of rating scale, which can give learners detailed information about their writing according to assessment categories (Hamp-Lyons, 1995).

In Chapter Five, English teachers at FLHSK responded in the questionnaire survey that they were not sure of the reliability and validity of subjective holistic scoring, especially the latter (see section 5.2.5 for details). In this chapter, I have discussed the characteristics and problems of subjective holistic scoring. Reliability was generally better than expected, except in the case of intra-rater reliability with Rater A, but the diaries and think-aloud data showed that the raters were likely to assess according to their hidden curriculum, which was not recognised by language testing researchers or compatible with the aim of the curriculum. The fact that subjective holistic scoring is problematic in terms of validity should not be overlooked when measuring the accountability of the guidelines, for invalid assessment through subjective holistic scoring does not help achieve the aim of the curriculum, which is to develop writing ability and contribute to national development and globalisation at macro level. If the course is assessed according to the teachers' subjective criteria, which are sometimes partial or differ from the guidelines, the assessment would not be accountable to the curriculum. Therefore, it can be concluded that subjective holistic scoring is not a scoring method to be recommended.

In light of this, I next asked the raters to assess the same scripts using one of existing rating scales, the FCE scale. The following sections will investigate whether the scale would be satisfactory in the Korean context.

## 6.3 Scoring using the FCE rating scale

I will begin by introducing the scale and procedure for the study, and then discuss the results of quantitative analysis of the ratings and qualitative analysis of the rating process using the scale.

### 6.3.1 Background to choice of the FCE rating scale

The FCE exam was "introduced by UCLES in 1939 as a preparatory examination to

CPE" and "originally known as the lower certificate in English" (Davies *et al.*, 1999: 62). Subsequently, it has been widely regarded as an examination for intermediate level. Therefore, given that the level of many Korean high school students may be around or below intermediate level, it could be a candidate rating scale for assessment of the English Writing course in question. Additionally, the guidelines to the 7th national education curriculum in Korea introduce the FCE scale for writing assessment as an example of rating scales. Therefore, I chose the FCE scale as a formal rating scale that contrasts with subjective holistic scoring for this study, and decided to investigate its appropriateness in the Korean situation.

### 6.3.2 The FCE scale for writing assessment

The FCE exam was introduced by UCLES in 1939, was revised several times after the Second World War, and was given its current, five-section form in 1996: Reading, Writing, Use of English, Listening and Speaking.

The Writing section has two parts. Part 1 consists of one compulsory task, to write a transactional letter, which may be formal or informal. Part 2 consists of four different task types, of which each candidate must choose one. The task types can be a composition, an article, a report, a letter of application, an informal letter and a short story. A composition is written for a teacher as a follow-up to a class activity, and may contain opinions and suggestions. An article is a type of description or anecdote written for a magazine or newsletter. A report is written for a peer group or a superior such as a boss or a teacher, containing facts, suggestions or recommendations. A letter of application is written to an individual or an organisation for a job or scholarship. An informal letter is written to a known reader in order to interest the reader, share an experience or explain feelings or opinions. A short story is written for a magazine or anthology, of which the typical reader might be a fellow-student or an enthusiast for a certain type of fiction. An example of task types in parts 1 and 2 of the Writing section can be found in Appendix 5.

A specific rating scheme is used to assess this writing section. It is a general framework to cover six assessment categories: 1) Content; 2) Accuracy; 3) Range of structures and vocabulary; 4) Organisation and Cohesion; 5) Appropriacy of Register and Format; and 6) Target Reader. The general rating scheme can be found in

Appendix 6.

These assessment categories are used across writing task types. However, since there are various types of writing task in the writing section of the FCE test, all the descriptors in each assessment category are devised to be task-specific, that is, the marking scheme focuses on criteria specific to each particular task. For example, a task-specific marking scheme for the letter is as follows:

Content

Major points: Letter must include all the points in the notes, i.e. say why writer can only travel in July, say which accommodation writer prefers and why, say which activities writer has chosen. Refer to ability and/or experience in activities (NB acceptable to write about one activity). Respond to the final question.

Minor points: Mention of two activities with ability and/or experience in them.

Organisation and Cohesion

Letter format, with early reference to why the person is writing. Clear organisation of points. Suitable opening and closing formulae.

Appropriacy of Register and Format

Formal letter.

Range

Language appropriate for asking for and giving information. Vocabulary relevant to the information given and asked for.

Target Reader

Would have enough information and details relating to the writer's stay

(FCE Handbook, 2001: 22)

On the other hand, the task-specific marking scheme for a report is specified as follows:

Content

Report should give factual information about which lessons and/or other activities should be filmed. It should also give clear reasons for the choice of these lessons and/or activities.

Range

Language appropriate to giving information and explaining. Vocabulary relating to school lessons and activities.

Organisation and Cohesion

Report should be clearly organised. Sub-headings an advantage. Introduction and conclusion.

Appropriacy of Register and Format

Register could range from the neutral to the formal, but must be consistent throughout. Formal report layout not essential.

Target Reader

Would know which lessons and other activities writer thought should be filmed and why.

(FCE Handbook, 2001: 24)

Thus, the FCE rating scheme specifies the satisfactory requirements of each assessment category according to the type of task. These specifications of task types should be referred to in conjunction with the general marking scheme (see Appendix 6).

### 6.3.3 The scoring procedure

For the sake of convenience I made some changes to the FCE general rating scheme before the three raters started to use it. One change concerned the numbering of bands, and the other was the form taken by the rating scheme.

First, I changed the numbering of the bands from 0-5 to 1-6, for consistency across the three rating schemes, that is, subjective holistic scoring, the FCE scoring and the RS1/RS2 scoring, given that the first one had been done on a scale of 1-to-6.

Second, I modified the rating scheme from a holistic scale to an analytic scale. As can be seen in Appendix 6, the rating scheme substantially functions as an analytic scale including six separate and specific assessment categories (Content, Accuracy, Range, Organisation and Cohesion, Appropriacy of Register and Format, and Target Reader), although it is ostensibly in the form of a holistic scale. However, a holistic scale cannot reveal how the raters think of and interpret each assessment category. This is supported by Hamp-Lyons (1995) as follows:

> A holistic scoring system is a closed system, offering no windows through which teachers can look in and no access points through which researchers can enter (p. 760-1).

I believed that an analytic scale would be more helpful in investigating how the raters interpret each assessment feature, as it would lead them to assess and mention each category individually in their diary entries and think-aloud, which would help show how they understand the scale. Although I modified the form of the scale, there was no change in the descriptors for each band, so the content was maintained.[52] The modified rating scheme, which was used for this study, is outlined in Appendix 7.

After modifying it, I introduced the scale to the raters, who had neither seen nor used the FCE rating scheme. Based on the FCE Handbook (2001), I produced a

---

[52] This modification produced an analytic scale containing six assessment categories. This tends to be seen as a lot of categories, which may have made the scale seen more complicated.

leaflet which included an introduction to the test, the FCE rating scheme, the conversion table from total sum of grades from each assessment category into a single number,[53] and examples of rating. After having a look at the leaflet, the raters were asked to assess five scripts using the rating scheme and to keep a diary after the assessment. Next, I chose some of their most informative diary entries to use as reference points or models. The principle to keep a diary was the same as in subjective holistic scoring, but I intended to confirm, using the model diaries, that the raters needed to cover all of the six assessment categories in their assessment and diaries.

After this training stage, as with the subjective holistic scoring process, I asked the three raters to use the FCE scale to rate their assigned scripts on a band of 1-6 (Group A for Rater A, Group B for Rater B and Group C for Rater C). They were also asked to keep a diary of the assessment process, just as with the subjective holistic scoring, after referring to the chosen diary examples.

At least two weeks after rating the scripts using the FCE rating scheme, all the raters were invited to assess Group D and one of Groups E, F and G, and to do a think-aloud for the six specific scripts of Group D (see Table 5.1 in section 5.6.4). I gave general instructions on how to do think-aloud during the FCE scoring, but did not prepare separate models of think-aloud for the FCE scoring, since the model for the subjective holistic scoring verbal protocol could be consulted. I asked the raters to refer to the models of think-aloud used for subjective holistic scoring to remind themselves how to do it. They recorded the think-aloud on the cassette tape that I sent them.

As a result of this exercise I obtained three hundred and ninety ratings, two hundred and twenty-four diary entries and eighteen recorded think-aloud from the FCE assessment. I transcribed the think-aloud protocols and analysed the obtained data. I will discuss the analysis results in the following sections.

### 6.3.4 Quantitative analysis

As in the case of subjective holistic scoring in section 6.2.2, I investigated intra-and

---

[53] Since it is generally not practical to report multiple scores for a script (Hamp-Lyons & Prochnow, 1991), it was decided to report a final grade in a single number for this study.

inter-rater reliability on the basis of the same rationale as in the case of subjective holistic scoring. All procedure for this were the same as those described in section 6.2.2, except that the raters used the FCE scale. To begin with, the analysis results of intra-rater reliability are presented in Table 6.3 below.

Table 6.3 Intra-rater reliability using the FCE rating scale

| | Rater A1 * Rater A2 | Rater B1 * Rater B2 | Rater C1 * Rater C2 |
|---|---|---|---|
| Intra-rater reliability | .356 | .873** | .839** |
| N | 20 | 20 | 20 |

As we can see, the correlation coefficients were high and significant at the level of .01 for Raters B and C, but not for Rater A. This implies that while Raters B and C consistently interpreted and applied the FCE scale (setting aside the issue of whether their interpretation was right or wrong), Rater A was not consistent when using the FCE scale, and that his intra-rater reliability was below significance level. As he went on assessing using the scale, his interpretation may have become better and better, or he may have become more and more confused. It was supposed that either pattern resulted in low consistency and this non-significancy.

However, the inter-rater reliability of the FCE rating is fairly high for all of the raters, as shown in Table 6.4. Noticeably, those which are associated with Rater C (between Raters A and C and between Raters B and C) are lower than those between Raters A and B. This implies that even though Rater C's intra-rater reliability was very high, his way of interpreting and applying the FCE scale was likely to be different from those of the other two raters. In addition, as in subjective holistic scoring, the inter-rater reliability involving Rater A, whose intra-rater reliability is not significant is also quite high and significant at the level of .01, exceeding expectation as in subjective holistic scoring.

Table 6.4 Inter-rater reliability using the FCE rating scale

| | Rater A | Rater B | Rater C |
|---|---|---|---|
| Rater A | 1.000 | .729** | .665** |
| Rater B | .729** | 1.000 | .665** |
| Rater C | .665** | .665** | 1.000 |

In addition, intra-rater reliability was higher when using the FCE rating scale than when doing subjective holistic scoring. There was also a slight increase in inter-rater reliability when using the FCE rating scale. This suggests that the use of a formal rating scale may help raise both intra-and inter-rater reliability above the levels achieved with subjective holistic scoring. This increased reliability as content-related evidence of validity implies that the use of a formal rating scale may be positive for validity. This should be empirical evidence of the advantage of using a formal rating scale over subjective holistic scoring. In the next section I will examine the raters' interpretation and application of the scale through qualitative studies.

### 6.3.5 Qualitative analysis: diary and think-aloud analysis

#### 6.3.5.1 Data analysis procedure

The same method was used as for subjective holistic scoring. More details of this method can be found in section 6.2.3.1 above.

#### 6.3.5.2 Overview of the analysis results

The analysis results are again presented in two parts: the observed patterns and tendencies of the FCE scale, and its unsatisfactory aspects, which are further divided into three headings: (1) unclear concepts within assessment categories; (2) inappropriateness of assessment categories; (3) inappropriateness of descriptors. Given that the FCE rating scheme was chosen for this study as a well-regarded example of published rating scales, these features might have been expected to be applicable to other published rating scales in general, rather than only the FCE scheme. They will be discussed one by one in the following subsections.

#### 6.3.5.3 Observed patterns and tendencies in use of the FCE rating scheme

Certain patterns and tendencies were observed when the raters assessed using the FCE scheme, as was the case when they assessed according to their own subjective criteria. First, even though they were using a formal rating scale, the raters brought in criteria of their own that they had employed in subjective holistic scoring. This can be seen especially from Rater A's following diaries and a verbal protocol:

ID: 4-A0409A
Band 4
The most noticeable characteristic of this script is that there is no paragraphing.....(*the rest of this entry is omitted*)

ID: 4-A0411A
Band 4
When I had a closer look at this script, I found that as the argument proceeded, the points in the latter part got shorter and shorter. The first two paragraphs look good, but the third is shorter and not developed enough, and the fourth is even worse. These last two paragraphs don't look sufficiently developed. I suppose it might be because the writer wanted to finish more quickly, but it has meant that the writer did not maintain a balance between paragraphs.

ID: A-FCE-TA-01
Band 3
(*The rater starts reading the whole script*)
Hi, ____. I heard you'll visit Korea next week.
....(*the middle of the script is omitted*)
Bye~
Um..... **First of all, this script is too short**.... Well.... Although it is short, the content seems quite good... mmm... maybe there's not enough content... Well.... as for accuracy,.. well... it looks fairly accurate, but....there are errors in the superlative, for example.... Um.... For vocabulary.... because the script is so short the vocabulary doesn't seem very rich.... Next.... As for organisation and cohesion....well.... there's nothing that could be rated as good.... And... format.... except *hi* in opening, there are... um... no words or expressions indicating that this is a letter. Um... therefore... it looks like this writing is not interesting.... Um... Content, Accuracy and Vocabulary could match Band 3... because it is too short, organisation and cohesion is insufficient... and it is not interesting... As for the other categories, Band 2 seems right.

As noted in section 6.2.3.2, Rater A gave great weight to length, paragraphing and development of argument in subjective holistic scoring, and even included these when using the FCE scheme, which does not include these criteria explicitly.

Second, although the raters were using the rating scale, they sometimes assigned or wanted to assign a band based on their impression from a first reading, as demonstrated in the protocol and diary below.

ID: C-FCE-TA-02
Band 4
(*The rater starts reading the whole script*)
To _____,
Hi, _____. How have you been?
.....(*the middle part of the script is omitted*)
See you in this winter. Take care.

With regard to Content,.... it would be Band 5 because this script deals with two places, such as Seoul and Jeju-do, not completely but fairly.....
As for Accuracy....there are too many grammatical errors and spelling errors.... Um..... how about Band 3?
As for Range.... well.... good idioms... long and good sentences..... so Band 5?
Organisation and cohesion..... which means a close relationship between sentences.... in this aspect, this script does not look complete... so it can be assigned Band 4...
In terms of Format and register, on the whole this script doesn't conform to their rules .... So could be Band 4...
As for Target Reader..... which means whether this script is appropriate for the readers.... it has many errors in this aspect too.... I would like to give Band 4....
**This script is very long and a lot of effort has gone into it, so I wish I could give higher bands... but unfortunately I can't mark it any higher than this ...**

ID: 4-A0408A
Band 4
Upon closer examination of this script, I found it to be at a low level, primarily due to quite a number of grammatical errors. On the other hand, this script looks as if it should get very high bands in terms of the other assessment categories apart from Accuracy. To be honest, I didn't want this script to be assigned a high band after all, because the grammatical errors didn't warrant it. So in order to lower the total band I assigned Band 2 for Accuracy. I admit I was too strict.

Finally, as with subjective holistic scoring, the raters were sometimes affected by the level of the preceding script. This is revealed in the following diary entries:

ID: 4-A0433A
Band 4
It occurs to me that when I assess, one of the factors affecting how I grade a script is the time gap between assessing one script and the next. I mean, I wonder how much time I need to have between assessing two scripts in order to prevent my assessment of the first from affecting how I rate the second. I have come to think about this because I do feel that the rating given to one script affects that of the next one. I compare the two scripts and fail to assess either of them objectively. Of course I know that a human is not divine, and it is impossible to assess every script objectively. What I would like to say is that we need to think about how much time should be left between two assessments to ensure as much objectivity as possible. Anyway... this script is well paragraphed, unlike the previous one.

ID: 4-A0426A
Band 5
**This script is written more clearly than the previous one.** The writer's thought is developed both smoothly and clearly, and its organisation looks good.

To summarise, certain patterns and tendencies emerged when the raters were

assessing using the FCE rating scheme. Although they were using a formal rating scheme, they were sometimes influenced by their own subjective impression and the order variable.

### 6.3.5.4. The problems with the FCE rating scheme

#### 6.3.5.4.1 Unclear concepts within assessment categories

There are six assessment categories in the FCE rating scheme: Content, Accuracy, Range, Organisation and Cohesion, Appropriacy of Register and Format and Target Reader. Their diary entries show that the raters felt unsure about some of these categories, and appeared to have difficulty understanding and applying them. Appropriacy of Register and Format was a case in point, as the following diary entries and a verbal protocol reveal:

> ID: 4-B0135B
> Band 4
> ….(*the first part is omitted*) The most peculiar error in this script is that it is not well-paragraphed and is written in a very colloquial style. Since this is the case, I suppose that marks will have to be cut in terms of both Accuracy and Register, but I am not sure. Still, I am unsure about the category of Register.

> ID: 4-A0408A
> Band 4
> ….(*the first part of this diary is omitted*) I feel that the category of Register and Format is getting more and more ambiguous as I assess scripts using this FCE rating scheme. I suppose that if a script is about the advantages and disadvantages of the Internet, due to the formality of its content, the script naturally becomes formal…(*the rest is omitted*)

> ID: B-FCE-TA-05
> Band 4
> (*The rater starts reading the whole script*)
> Today, many people surf the internet at leisure time. Internet is very popular these days among teenagers as well as adults…..(*the middle of the opening part is omitted*)….Like all other human invention, Internet has not only some advantages but also some disadvantages.
> Um…. up to this point it looks like this is the opening part of this script… I wonder why the writer starts every sentence in a new paragraph…
> Now, let's look at some advantages first.
> Here it seems to enter the body. But this paragraph is also composed of just one sentence…..
> Above all, by using the Internet, we can collect a lot of information and share them easily. For example, if you want to do a homework about 'Emile Jong', you can surf the Internet site such as 'Naver' and search for information about 'Emile Jong' very quickly and easily. So you can do your homework in about 30 minutes.

Second, you can communicate with other people all over the world by using the Internet.
But Internet has a lot of disadvantages, too.
The writer is discussing two advantages, but one of them is long and the other is very short with just one sentence..... And the paragraph which introduces the disadvantages consists of just one sentence, so the paragraphs don't seem to be developed equally.
First of all, you can be addicted to Internet if you use the Internet too long. There are plenty of people who are addicted to internet on-line game or lewdness site in every PC Room. They surf the Internet all day long and don't go to school or work. Such people need psychological treatment, but these people are increasing gradually.
Second, we approach bad information such as lewdness site or violence site, suicide site or bomb site. .... I wonder if this point is similar to the first one?... Therefore, we should use the Internet wisely so that the Internet can contribute to our society.
.....(*the middle of this protocol is omitted*) And next... **as for Register and Format,... it is also .... well... Band 4... I am really unsure what to assess in this category**....(*the rest of this protocol is omitted*).

In addition to Register and Format, they were also unsure about the category of Range, as the following diaries show:

ID: 3-B0111B
Band 4
...(*the first part of this diary entry is omitted*) To be honest, I'm not sure about how to assess the Range of scripts. It might be partly because I haven't fully grasped the concept of Range, and partly because there is little difference between Korean high school students' scripts in terms of vocabulary and structure. To me most scripts look similar in this category....(*the rest of this diary entry is omitted*).

ID: 4-A0424A
Band 4
I had to read this script a few times before assigning a band, which means it is hard to understand. I suppose this is because it is not logical and includes some very inappropriate words which make it hard to grasp the meaning. I am not sure which aspect I should mark it down in because it is not written logically. Is it in Content or Range? I did Content and assigned Band 4 for that, and I gave the same band for Range for the use of inappropriate words. Consequently, I came to assign the same bands for both Content and Range again.

Target Reader was another category that the raters were unclear about.

ID: 4-B0127B
Band 4
This script is well organised in that it consists of opening, body and closing, and in that each of three points in the body is developed equally in three paragraphs. Unfortunately, although the required content from the prompt was to discuss the

advantages and disadvantages of the Internet, this script discusses only the disadvantages of the Internet. It seems to me, therefore, that this script does not satisfy the requirement of Content. Apart from this problem, the content of this script doesn't look good. So I wonder how much I should mark it down in terms of Content. Further, given that half of the major content is omitted, I suppose that this script only achieves half of its purpose, so there is good reason to assign a low band in terms of Target Reader as well. **But I am not sure whether I am right and whether I understand the concept of Target Reader**..... (*the rest of this diary entry is omitted*).

These difficulties were probably due to the fact that the concepts of Register and Target Reader were unfamiliar to the raters as they were never included in their subjective criteria, and thus they may have felt uncomfortable with such notions when using the FCE rating scheme. Additionally, the descriptors were not clearly worded or given in detail. For example, the category of Organisation is described like this in Band 5: "Ideas effectively organised". The category of Target reader is described for Band 5 with "Fully achieves the desired effect on the target reader". These descriptors are so vague that the raters had good reason to be unsure of the concepts, and could not help but project their own assumptions on. As for Target Reader, for example, Rater A understood this as the degree of interest a script would hold, Rater B as both the degree of interest and the degree to which the script achieved its purpose, whilst Rater C viewed it as the degree of communication with the reader, as can be seen in the following diary entries:

ID: 3-A0436A
Band 4
...(*the first part of this diary entry is omitted*) I also have trouble assessing Target Reader, which is meant to determine whether a script is interesting. I wonder if this writing dealing with a myth is interesting. It seems to depend on the rater's viewpoint. I just decided on Band 4 for this because I found this script kind of interesting....(*the rest of this diary entry is omitted*)

ID: 3-B0101B
Band 4
....(*the first part of this diary entry is omitted*) When it comes to Target Reader, while it is not particularly interesting to read, it is fairly successful in achieving its aimed purpose of introducing some places to visit. Therefore, I decided Band 4 for this category...(*the rest of this diary entry is omitted*).

ID: 3-B1115C
Band 4

... *(the first part of this diary entry is omitted)* For Target Reader, since this script communicates well on the whole, I decided on Band 4 for this script....(*the rest of this diary entry is omitted*).

A similar phenomenon was observed regarding Organisation, as can be seen in the following diaries and a verbal protocol:

ID: 3-A0407A
Band 4
...(*the first part of this diary entry is omitted*) For Organisation, it is not paragraphed, nor does it conform to the genre format of letter. For Cohesion, this writing doesn't seem to include any necessary linking devices..... (*the rest of this diary entry is omitted*)

ID: 3-B0102B
Band 3
...(*the first part of this diary entry is omitted*) This script follows the genre format of explanatory essay rather than the genre format of letter required from the prompt. It is not organised into opening, body and closing either···(*the rest of this diary entry is omitted).*

ID: 3-B1130C
Band 5
....(*the first part of this diary entry is omitted*) This script does not seem to be well organised because it introduces only one place, Bongeun temple, rather than various places. Therefore, I assigned Band 5 to this writing for Organisation and Cohesion.....(*the rest of this diary entry is omitted*).

ID: A-FCE-TA-03
Band 3
(*The rater starts reading the whole script*)
Hi, I'm going to recommend you a great place to visit in Korea. If you search in the map, you'll find a big island in very south of Korea. It called Jeju island....(*the middle of the script is omitted*)... I hope to see you in Jeju 2 years later. I'll write to you more about Jeju later.
.....(*the middle of the script is omitted*)
Bye. With love, _____.
This script.... um... seems to have a relatively good format, but.... well...the Content seems.... well... middle level.... Say.. Band 3... Grammatical accuracy seems to be very poor... There seem to be errors which are "critically" inaccurate rather than inaccurate "on the whole"... well...how about Band 2 for Accuracy? Right... Band 2 would be appropriate... Next, Range... it looks like a similar level to Content.. So Band 3 would be fine.. **Next, pertaining to Organisation and cohesion, although it is paragraphed, it is not done well ... um.. as it is done anyway, Band 3 or so would be fine...**(*the rest of the protocol is omitted*).

#### 6.3.5.4.2 Inappropriateness of assessment categories

The second problem with the FCE scheme, which might make it unsuitable for the Korean situation, is the inappropriateness of assessment categories. This can be demonstrated in two ways. First, they mentioned that it included unnecessary categories and lacked necessary ones. They found some assessment categories in the FCE rating scheme, such as Register, unnecessary in the Korean situation. They thought the categories were inappropriate because they did not help differentiate students' writing, as revealed in the following diary extract.

> ID: 4-B0103B
> Band 5
> ... (*the first part of this diary entry is omitted*) Looking back on the three previous scripts, I have to admit that I assessed the category of Register without really understanding it. I wonder if there are differences between Korean students' scripts in terms of Register. I suppose they just pay attention to "making sentences" that are accurate and grammatical because of their intermediate or low level of English, so they won't have much variety of register depending on the situation given in the prompts.
> So I don't think there is much difference between them in terms of Register. If this is the case, I don't think that this category is necessary in the Korean situation....(*the rest of this diary entry is omitted*).

In contrast to this, Rater A mentioned that there were some categories that needed to be added to the scheme. One rater wondered how 'length' could be assessed and reflected in the rating scheme, as shown in the following diary:

> ID: 4-A0428A
> Band 3
> I had trouble assigning bands for all of the assessment categories in the rating scheme. This was because this script was too short to apply to the rating scheme. I think the descriptors in the rating scheme assume that the script is longer than the minimum length. So the descriptors for this rating scheme do not look appropriate for this kind of short script.
> As this is too short, I assigned Band 2 for Content, since it did not even include minimum content. Pertaining to Accuracy, there are few errors. Nonetheless, I would not like to assign a good grade, because the lack of errors is not due to its excellence but to its shortness. However, there is no descriptor for this situation in this rating scheme. I had trouble assessing Organisation and cohesion, too. Although I know that Organisation and cohesion do not have to do with length, I didn't want to give a good grade because each paragraph is too short. There is no band to deal with this situation, either. As for Cohesion, the use of linking devices looks appropriate in its own right, but I didn't want to give a good grade for this as there are very few linking devices in the script. I don't think it is fair to give good grades just because one or two linking devices are used appropriately....

To prevent this short script from getting good grades, I think there should be a category or descriptors to deal with this situation.

Rater A also noted in his diary that it was desirable to add the category of 'development of idea' to the FCE rating scheme:

ID: 4-A0433A
Band 4
.... (*the first part of this diary entry is omitted*) This script is paragraphed, but it is not done appropriately. Each paragraph is too short and not developed enough. However, it looks as though there is no descriptor to deal with this kind of situation.... (*the rest of this diary entry is omitted*)

The raters also mention a category dealing with awkward/Korean-like expressions (Konglish). When examining the scripts of Korean high school students in this study, we can see that their English appears to be affected by the Korean language due to language transfer/cross-linguistic influence (Benson, 2002),[54] so their writing sometimes appears to be directly translated from Korean into English. Put another way, they sometimes use Korean-like English, or Konglish. This Konglish is awkward, and negatively affects the rater's impression of the quality of writing. Although it does not always make a sentence ungrammatical, it is not an aspect that the raters are willing to overlook, but one which they are likely to mark down for. Since there was no category to deal with this in the FCE rating scheme, the raters felt uncomfortable with it, as revealed in the diaries and a verbal protocol.

ID: 4-A0421A
Band 4
I had trouble with awkward expressions. They are not ungrammatical, just awkward. In this case, I don't think it's fair to mark it down for Accuracy because they are not inaccurate, but as there is no category to deal with this, I cut the marks for Content. However, I know this application was still not right..... (*the rest of this diary entry is omitted*).

ID: 4-B0138B
Band 4

---

[54] According to Benson (2002), this is probably the case. She writes that when L2 learning takes place in classrooms which lack opportunities for authentic input and interaction, cross-linguistic influence (or transfer) is more likely to take place than in naturalistic settings. Given that their writing activity took place in classrooms and, furthermore, that they have generally learned English in that setting, cross-linguistic influence seems likely.

This script includes many awkward expressions which cannot be said to be grammatical errors. These seem to be constructed unnaturally, translated literally from Korean into English. I wonder, however, which category deals with this aspect in this rating scheme.

ID: B-FCE-TA-04
Band: 5
(*The rater starts reading the whole script*)
Nowadays, it's very hard to find our homes without internet connection. Almost every apartment and houses are using ADSL or Cable, called High-speed internet. Korea internet use is highest percentage in the world. Now we can find how much our lives depend on internet. Then we must know about strengths and disadvantages of internet. Let's think about them. Good opening...
Internet makes our lives 'comfortable'. We can find useful information so easily, like music, studies, movies, news, maps and so on. Just type what you want to find and click the 'search' button. Then maybe..... wanted information appear on the screen.
Almost internet boards are useful to share our opinions. We can discuss and talk about what we are interested in, not concerned by..... where they live, what they do, and what they appear to be. This sentence looks awkward... seems to be long... and Konglish...
Last, we can 'share'. Things that few rich or power-gained people had in the past are shared by everybody now. Some sharing is illegal, but it can contribute to our society and develop.. This should be in the form of noun, *development* rather than verb type...of democracy.
However, as internet-using people increases, many problems broke out. One thing is internet addict. Many people surf and play games on the internet or chat. They feel uncomfortable when they cannot use internet. Some people cannot distinguish the online and the offline.
And, internet is communication of all the people Among... who.... don't know each other... This sentence looks wrong..., starting with relative pronoun... . why is *among* used here?... in the reality. Some people use other's or.... *Other's*? .. possessive form is not required here... do not use their names and criticize others without evidence or any logical reason. That may be a big problem.
As we write email instead of handwritten mail, we forget the importance of analog. We must know digital is not the whole that can give us bright future. We should combine digital and analog, natural and artificial things.
As Korea has the project of being top. .....it looks unnatural...It industry country, all the citizens must know about how to make good use of internet. Then we can be improved, life standards and the nation will be, too.
.....For Accuracy,...although there are some grammatical errors..... relatively good on the whole... however, there are some sentences which are not inaccurate but awkward and Konglish.... I wonder what band is suitable for this case....probably Band 5?...since there are not only some inaccurate sentences but also awkward and Konglish sentences, I am not sure how I should apply the rating scheme.....(*the rest of this protocol is omitted*).

Second, according to their diary entries, the assessment categories which include two assessment features, such as 'Organisation and cohesion' or

'Appropriacy of Register and Format' would be more useful if they were separated into two separate categories. This is because the raters tended not to assess both subcategories in such a composite category, but were likely to focus on one or the other. This problem is shown in the diary entries below:

> ID: 4-B1125C
> Band 5
> ...(*the first part of this diary entry is omitted*) In terms of Organisation and cohesion, this script could be assigned Band 6, but since I think it is less organised in that advantages and disadvantages of the Internet are not fully discussed, Band 5 will be the maximum band for this script. For Appropriacy of Register and Format, since it seems that this script conforms to the format for formal essays, Band 5 looks appropriate for it..... (*the rest of this diary entry is omitted*).

> ID: 4-B1117C
> Band 6
> ...(*the first part of this diary entry is omitted*) Since the linking between sentences is good, I assigned Band 6 for Organisation and cohesion. On the whole this script is fairly good for Appropriacy of Register and Format, but given that the writer uses very pedantic expressions such as "*pros and cons*", instead of "*advantages and disadvantages*", which does not seem to be appropriate for learners, I chose Band 5 for this category.... (*the rest of this dairy entry is omitted*)

As can be seen from the diaries above, the raters tended to determine a band for the composite category on the basis of one of the two subcategories in it.

#### 6.3.5.4.3 Problems with descriptors of the FCE rating scheme

Problems were found not only with assessment categories, as mentioned in the previous two sections, but also with descriptors. First, the differentiation between bands mainly relies on either the use of quantifiers and degree adverbs such as *all, some, little* and *limited,* or ambiguous words between which the differences are not clear, such as *effectively, clearly* and *inadequately*. Consequently, the raters had difficulty grasping the differences between bands, as shown in the following diary entry and one verbal protocol:

> ID: 3-B0109B
> Band 4
> This script introduces Ha-hoe town in Ahndong in detail. Unfortunately, it has quite a number of grammatical errors. The errors are not local but global errors, which affects my understanding of what the sentences mean. For assessment of this, I had a look at the descriptors in Accuracy, hoping to find the most appropriate band for

this case. However, the words in the descriptors, such as *a number of errors* and *frequent errors* look very ambiguous to me. I cannot see the difference between them and I am not sure which would be more appropriate for this situation. As neither of them are clear to me, I just chose Band 3, according to my intuition.

ID: B-FCE-TA-01
Band 4
(*The rater starts reading the whole script*)
Hi, ____. I heard you'll visit Korea next week.
..... (*the middle of the script is omitted*) Bye.
...(*the first part of this protocol for other assessment categories is omitted*) Next, for Appropriacy of Register and Format.... the concept of this category is still ambiguous to me.... Anyway, this script is written neutrally and doesn't specifically reflect the genre of letter. So is it "reasonable" enough to be assigned Band 4, or "inconsistent" enough to be assigned Band 3? What on earth is the difference between "reasonable" and "inconsistent"? I just can't grasp the differences between the bands in this category, but... since this script is sort of good to read anyway, I'll choose Band 4. Next.... (*the rest of this protocol is omitted*)

Because the distinction was only made on the basis of ambiguous adverbs, the raters relied on subjective judgement for the band differences, as seen in the following diary:

ID: 3-B0109B
Band 4
This script introduces Ha-hoe town in Ahndong in detail. Unfortunately, it has quite a number of grammatical errors. The errors are not local but global errors, which affects my understanding of what the sentences mean. For assessment of this, I had a look at the descriptors in Accuracy, hoping to find the most appropriate band for this case. However, the words in the descriptors, such as *a number of errors* and *frequent errors* look very ambiguous to me. **I cannot see the difference between them and I am not sure which would be more appropriate for this situation. As neither of them are clear to me, I just chose Band 3, according to my intuition.**

It is not easy to completely avoid such ambiguous words in rating scales. However, it may be more helpful if the differences between the bands in a rating scheme rely not only on the kinds of words that indicate quantity, but also on qualitative differences, such as which features occur in a band.

Second, the FCE rating scheme is specifically designed for the FCE writing assessment, which has a specific type of questions (see section 6.3.2). Accordingly, the rating scheme seems inappropriate to different types of writing assessment.

Descriptors such as *major content* and *minor content* for Content in the FCE rating scheme do not appear appropriate when the scheme is applied in different assessment situations, including the type of writing assessment in this study. This problem can be seen in the following diary entry and think-aloud protocol:

> ID: 3-B0101B
> Band 4
> ...(*the first part of this diary entry is omitted*) For Content, this script introduces only one place, Jeju island. Although this script introduces only Jeju island, it still meets the requirement in the prompt to introduce places to visit in Korea. In this case, I am not sure whether this achieves the criteria of Content dealt with in terms of major content and minor content. That is, I am unsure whether this script covers major content only and omits minor content, or whether it does not even cover major content because it introduces only one place. After having trouble with this, I assigned Band 4, because obviously including only one place cannot be said to be enough ....(*the rest of this diary entry is omitted*)

> ID: B-FCE-TA-01
> Band 4
> (*The rater starts reading the whole script*)
> Hi, ____. I heard you'll visit Korea next week.
> ..... (*the middle of the script is omitted*)
> Bye~
> Um....to begin with, for Content, this only introduces Jeju island .... but the writer tried to introduce it in detail... I am not sure whether this includes major content but not minor content... I find this sort of case a bit tricky ....let's have a look at the descriptors in Bands 3,4,and 5 for Content,... Band 3 is the level where major content is inappropriate or omitted... But this script only introduces one place in detail, Jeju island .... given that there is no instruction in the prompt about how many major contents should be included, how can I assess this case? And how can I assess writing which includes some places, compared with this case?... I'm not sure about this point... anyway, because this script only deals with one place it can be said to omit major content, so I think Band 3 would be appropriate ....

Therefore, the FCE rating scheme is inappropriate for assessment of the English Writing course in Korea, which does not include the same type of writing assessment as the FCE writing assessment.

Finally, the descriptors of the FCE rating scheme appear very logical and hierarchical, but they do not include features that occur frequently in Korean students' scripts. This is the case with the category of Range, as seen in the next diary entry. According to the rating scheme, the script deserved a high level because the student attempted a range of vocabulary, but the rater hesitated to assign a good mark because

the use of vocabulary looked inappropriate and awkward. The rating scheme descends from a level that uses a range of vocabulary to a level that uses very limited vocabulary, but there is no appropriate descriptor in the FCE rating scheme for cases where a range of vocabulary is tried but is not appropriate.

> ID: 3-B0109B
> Band 4
> … (*the first part of this diary entry is omitted*) As for Range, it looks like this writer tried to use a variety of words, but their uses are awkward or inappropriate. In this case, I am unsure what band to assign to this script. The rating scheme does not address this situation. Having trouble with this point, I just chose Band 4…

In all the cases discussed above, the raters came across problems when using the FCE rating scale, and when they experienced these problems, they inevitably relied on their subjective impression to assign a band. Many of the problems were caused by the fact that the scale was not designed specifically for the Korean students and writing assessment in Korea.

### 6.3.6 Conclusions

I have pointed out the observed patterns, tendencies and problems in applying the FCE rating scheme to the assessment of Korean students' writing, revealed through quantitative and qualitative analyses. From the quantitative analysis, it was found that the intra- and inter-rater reliability was higher with the FCE scale than with subjective holistic scoring, although Rater A's intra-rater reliability was still not significant. This suggests that using a formal rating scale may be more conducive to reliability than subjective holistic scoring.

From the qualitative analysis, it was found that the raters were sometimes affected by their own subjective judgement and by the ratings of preceding scripts. The problems with the FCE rating scale were that some assessment categories were both unclear to the raters and inappropriate for assessing Korean students' scripts, causing problems with the validity and thus the accountability measure of this FCE scoring to the aim of the national curriculum, as in subjective holistic scoring, and that some descriptors in the scale were also problematic because they relied solely on

quantifiers[55], were specifically for the FCE writing test and did not reflect the characteristics of writing by Korean students.

In sum, the problems with the FCE rating scheme discussed so far in this study are as follows: the scale is not fully satisfactory in its own right; it was devised for a different test; it was not specifically devised for Korean students. Therefore, it does not appear to be satisfactory to use the FCE rating scheme to assess Korean students' writing or for the English Writing course.

## 6.4 Summary and conclusions

This chapter sought to investigate whether subjective holistic scoring and the use of the FCE rating scale, which was one of three rating schemes available to Korean English teachers to assess their students' writing performance,[56] were appropriate for assessment in the given context. In addition to arguing on the basis of previous studies and the findings from the Questionnaire I survey in Chapters One and Five, in the present chapter I sought to verify my judgement that these two rating schemes were problematic in terms of both reliability and validity, and to determine whether it would be desirable to develop a new rating scale for this context.

To this end, the three raters were asked to assess the obtained writing samples, first by a subjective holistic scoring method and then with the FCE rating scale, and to keep a diary and to do a think-aloud in order to illuminate their rating process.

Intra- and inter-rater reliability were high in the rating through subjective holistic scoring, with one exception of Rater A's intra-rater reliability, which suggested that this type of scoring is more reliable than expected.

In the diary and think-aloud studies, the following features were found: none of the raters missed assessing grammar and content; they all showed a central tendency

---

[55] Similar problems have been pointed out in previous studies. Matthews (1990) claims that the individual categories in the scale of the International General certificate of Secondary Education (IGCSE) are not always clearly defined, and that this creates problems of validity. Additionally, pointing out that band descriptors of the rating scale for ELTS are described in only vague general terms and abound in quantifiers, such as "*at times*", "*some*", "*not always*" and "*most of*", she contends that through these terms only gross distinctions can be made and that this affects the reliability of assessment.

[56] In Chapter One I suggested three rating schemes available to English teachers at FLHSK at present: subjective holistic scoring, using a rating scale developed by the teacher/teachers and using one of the published rating scales. I investigated the appropriateness of the first and the third schemes for the given context. The second scheme was not covered in this study, but could be the subject of a further study.

in rating; they paid attention to different assessment categories depending on the level of a script; they believed that scripts in the higher bands needed to include advanced structure, vocabulary and expressions; they assessed questionable aspects such as effort; they were affected by the ratings of preceding scripts; they hesitated between two bands before choosing one band for a script; they were not always confident of their rating; they did not establish clear criteria for the highest and lowest bands; and their ratings were arrived at through an unobservable and varying internal negotiation process. Given these findings, subjective holistic scoring appeared to be of questionable validity. Therefore, I moved onto the investigation of another rating scheme, using a published rating scale.

Of the many published rating scales available, I chose the FCE scale for writing assessment for this study. Since the FCE is for learners at intermediate level, its rating scale was considered a possible candidate for the assessment of Korean students. The same procedures were followed as for the investigation and analysis of subjective holistic scoring. In the quantitative study, higher reliability was achieved than with subjective holistic scoring, even though Rater A's intra-rater reliability was still very low. This implied that the use of a formal rating scale may have helped increase reliability. The observed patterns and problems in the qualitative study were specifically as follows: the raters were sometimes affected by their subjective holistic judgement; their rating was influenced by the ratings of preceding scripts; some assessment categories in the scale were unclear and inappropriate for assessing Korean students' writing; and there were problems with the descriptors in the scale – for example, descriptors included quantifiers, were specifically for the FCE writing test and did not reflect the features of Korean students' writing.

These results show that there are particular limitations associated with subjective holistic scoring and the FCE scale. Additionally, as discussed in Chapter One, the FCE scale was not devised to reflect the construct and goal of this specific course, the English Writing course at FLHSK. Therefore, there is a need for a new rating scale for this given context. In the next chapter I will discuss the development of the new scale.

In light of the literature reviewed in Chapters Two to Four and the context in question, the new rating scale should meet the following requirements: it needs to be

used mainly for free writing in direct assessments for the English Writing course in FLHSK; it needs to be in an analytic scale, given the benefits of this type of scale over a holistic scale (as discussed in section 4.4) and Korean students' need for detailed and diagnostic feedback on their writing (see section 5.5.1); the scale needs to be assessor-oriented to help Korean teachers rate, given that they have not been provided with any rating scales for the assessment of the course; it needs to be developed by a data-based approach, taking into account the defects of the *a priori* approach to rating scale development, as discussed in section 4.5.2; it needs to fit the objective of the course, which is the ability to write in English in an organised, accurate and fluent manner for effective communication, which, if expressed from theoretical and pedagogical perspectives as discussed in section 2.4, is the ability to produce both 'acontextually' and contextually correct forms of language following the prescribed patterns at both sentence level and at discourse level, so as to communicate with readers functionally. From the next chapter, I will discuss the attempt to develop such a scale.

# CHAPTER SEVEN. WRITING SAMPLE ANALYSIS AND THE DEVELOPMENT OF A RATING SCALE

## 7.1 Introduction

In the previous chapters I made my arguments for the need for a new rating scale for FLHSK, and at the end of Chapter Six I indicated what kind of scale needs to be developed. As noted above, I chose a data-based approach to develop this scale, and decided to construct it on the basis of writing samples from Korean students. In this chapter I will outline how the written corpus obtained was analysed before the rating scale was developed.

In section 7.2 I will introduce the general methodology followed for coding of the writing samples and the development of the rating scale, and in section 7.3 I will discuss the coding scheme that was constructed. In section 7.4 I will outline the coding procedure; in section 7.5, the procedure and results of statistical analysis of the coding will be examined to draw out features discriminating between neighbouring bands. Section 7.6 will cover the development and use of the RS1 on the basis of the features obtained from the statistical analysis. Section 7.7 will outline the revision process of the RS1 to produce the RS2. Finally, I will summarise this chapter.

## 7.2 Analysis methodology

There are several methodologies that can be used to analyse writing samples in a data-based approach to rating scale construction, as discussed in section 4.5.3. They can be divided into two main types. One relies on the judgement of teachers or rating experts to find specimen samples or criteria for a band (e.g., Alderson, 1991; Griffin, 1990; Upshur & Turner, 1995). The other analyses data using the researcher's own coding scheme (e.g., Fulcher, 1993, 1996b).

Elements of both methods were adopted for this study. To begin with, teachers' ratings from subjective holistic judgements in Phase One were employed in order to classify the samples into six-band groups. Then, all the written data were analysed according to a coding scheme that I developed for this study on the basis of the features of the data, the writing ability implicit in the course objectives and literature

review. I will introduce the coding scheme in detail in the next section.

## 7.3 Development of the coding scheme

### 7.3.1 Coding scheme development procedure

I decided to use three starting points to develop the coding scheme: the definition of writing ability in the English Writing course at FLHSK; the features of the obtained scripts; and theoretical considerations. First, for the main categories of the coding scheme, I believed that the main construct of writing abilities used in the course objectives needed to be reflected in the coding scheme for the concern of *a priori* construct validity of both the coding scheme and the RS1 that would be derived from the scheme. As mentioned in section 2.4, the main construct of writing abilities used in the course objectives is the ability to express one's own thoughts and feelings in an organised, accurate and fluent manner across various genres for effective communication. Therefore, the three main constructs in the objectives of the course, namely accuracy, fluency and organisation, were chosen as the main categories in the coding scheme. However, since these three categories were abstract, they needed to be operationalised for the actual analysis. To this end, the written data from the subjects were investigated before substantive categories for each of the three categories could be developed. Looking at the data, I tried to find the features in each of the main categories that could help differentiate between the bands. To this end, I repeatedly read the scripts, noting their salient features, and reviewed previous literature on such salient features. As a result, I devised a coding scheme with eighty-two categories. I will discuss this coding scheme in detail in the next section.

### 7.3.2 Coding scheme

Through the procedure outlined in section 7.3.1, the coding scheme that was to be used to analyse the scripts was developed as shown in Table 7.1 below. Each category is followed by a bracket that shows the coding method used when coding the writing samples for the category. For categories marked 'C', the number of the category occurrence in a given script is counted and for categories marked 'T', the coder makes a binary choice or ticks one subcategory in the category that best describes the script.

Table 7.1 The coding scheme for coding scripts

(1) Accuracy
  (1.1) 'Intelligible'
    (1.1.1) Verbs
      (1.1.1.1) Distinction between finite and non-finite verbs/omission or repetition of finite verbs (C)
      (1.1.1.2) Distinction between verb types (C)
    (1.1.2) Tense (C)
    (1.1.3) Indication of quantity in nouns and Articles (C)
    (1.1.4) Agreement (C)
    (1.1.5) Conjunctions and Relatives
      (1.1.5.1) Syntactic error (C)
      (1.1.5.2) Semantic error (C)
    (1.1.6) Distinctions between word classes
      (1.1.6.1) *because, there & for example* (C)
      (1.1.6.2) Other (C)
    (1.1.7) Voice and Participles (C)
    (1.1.8) Prepositions and Particles
      (1.1.8.1) Syntactic error (C)
      (1.1.8.2) Semantic error (C)
    (1.1.9) *To-* or bare infinitives and Gerund (C)
    (1.1.10) Auxiliaries
      (1.1.10.1) *Do*-support (C)
      (1.1.10.2) Other
        (1.1.10.2.1) Syntactic error (C)
        (1.1.10.2.2) Semantic error (C)
    (1.1.11) Spelling, Capitalisation and Punctuation
      (1.1.11.1) Punctuation between main clause and subordinate clause (C)
      (1.1.11.2) Spelling (C)
      (1.1.11.3) Other (C)
    (1.1.12) Vocabulary and Phrase
      (1.1.12.1) Word coinage (C)
      (1.1.12.2) Inappropriate word and phrase (C)
      (1.1.12.3) Words that are literally translated from Korean, or phrases that are either ungrammatical or literally translated from Korean (C)
    (1.1.13) Clauses that are either ungrammatical or literally translated from Korean (C)
    (1.1.14) Other grammatical errors
      (1.1.14.1) Possessive (C)
      (1.1.14.2) Word order in a phrase (C)
      (1.1.14.3) Omission of subject in a finite clause (C)
      (1.1.14.4) Other (C)
  (1.2) 'Unintelligible'
    (1.2.1) Errors in clause construction
      (1.2.1.1) Due to serious syntactic error, resulting in unintelligible clause (C)
      (1.2.1.2) Due to errors in clause construction, resulting in ambiguous and unclear clause (C)
    (1.2.2) Use of unintelligible vocabulary (C)
    (1.2.3) Other (C)
(2) Fluency
  (2.1) Quantity (T)
    (2.1.1) Less than 33% of (implicit) minimum quantity of around 200 words (66 words)

(2.1.2) Between 33% and 75% of (implicit) minimum quantity (66 to 150 words)
(2.1.3) Around 100% of (implicit) minimum quantity (200 words)
(2.1.4) More than 150% of (implicit) minimum quantity (more than 300 words)
(2.2) Coherence
(2.2.1) Disconnected and incoherent sentences for more than 50% of the whole script (T)
(2.2.2) Local lack of coherence
(2.2.2.1) For an individual sentence
(2.2.2.1.1) Due to insufficient language command (C)
(2.2.2.1.2) Due to its irrelevance to the previous sentence (C)
(2.2.2.1.3) Due to its unintelligibility (C)
(2.2.2.2) For more than two consecutive sentences
(2.2.2.2.1) Due to inappropriate alignment of the sentences (C)
(2.2.2.2.2) Due to their irrelevance to the previous sentence (C)
(2.2.2.2.3) Due to their unintelligibility (C)
(2.3) Cohesive devices
(2.3.1) Little repetitions using the substitution words (T)
(2.3.2) Smooth connection between sentences using cohesive devices well (T)
(2.3.3) Use of advanced connectors (T)
(2.3.4) Errors in the use of number and person of pronouns (T)
(2.4) Advanced language
(2.4.1) English-like vocabulary, phrase and lexical phrases (T)
(2.4.1.1) Uses them once or twice
(2.4.1.2) Uses them more than three times
(2.4.2) Good clause construction and good expansion of clauses through fluent use of adjective/adverbial clauses (T)
(2.4.2.1) Uses them once or twice
(2.4.2.2) Uses them across the script
(2.4.3) Advanced grammar
(2.4.3.1) Use of complex aspect (C)
(2.4.3.2) Use of relative adverbs (C)
(2.4.3.3) Use of *that*- clause for complement and subject (C)
(2.4.3.4) Other (C)
(2.4.4) Multi-word verb phrases (C)
(3) Organisation
(3.1) Paragraphing (T)
(3.1.1) No paragraphing
(3.1.2) Errors in paragraphing
(3.1.3) Exact paragraphing
(3.2) Genre format and development
(3.2.1) Opening
(3.2.1.1) Blurred distinction between opening and body or lack of opening (T)
(3.2.1.2) Genre format (T)
(3.2.1.2.1) Not follows the genre format
(3.2.1.2.2) Follows the genre format
(3.2.1.3) Development (quantity) (T)
(3.2.1.3.1) Less than two sentences
(3.2.1.3.2) Between two and three sentences
(3.2.1.3.3) More than three sentences
(3.2.2) Body
(3.2.2.1) Number of points (C)
(3.2.2.2) Number of insufficiently developed points (C)

(3.2.3) Closing
(3.2.3.1) No closing (T)
(3.2.3.2) Genre format (T)
(3.2.3.2.1) Not follows the genre format
(3.2.3.2.2) Follows the genre format
(3.2.3.3) Development (quantity) (T)
(3.2.3.3.1) Less than two sentences
(3.2.3.3.2) Between two and three sentences
(3.2.3.3.3) More than three sentences, but not rounded off
(3.2.3.3.4) More than three sentences and reasonably rounded off

(3.3) Topic address (T)
(3.3.1) The purpose / topic of the discourse is not explicitly addressed in the opening stage of the discourse
(3.3.2) The purpose / topic of the discourse is not signalled in the opening stage of the discourse
(3.3.3) The purpose / topic of the discourse is wrongly addressed in the opening stage of the discourse
(3.3.4) The purpose / topic of the discourse is partly addressed in the opening stage of the discourse and the required topic is fully dealt with in the discourse or vice versa.
(3.3.5) The purpose /topic of the discourse is appropriately addressed in the opening stage of the discourse.

(3.4) Content (T)
(3.4.1) The discourse omits some of the required content from the prompt.
(3.4.2) The discourse includes irrelevant points to the topic.
(3.4.3) The discourse includes required content, but extremely simply and insufficiently.
(3.4.4) The discourse sufficiently includes required content.

Most of the above categories should be fairly self-explanatory; those which are not, or seem to require comment, are explained below.

#### 7.3.2.1 Accuracy

The category of Accuracy here is concerned with correctness and grammaticality at sentence (strictly speaking, clause) level, given that Richards, Platt and Platt (1992) define accuracy as the ability to produce grammatically correct sentences.

For coding the observed features in the scripts, as mentioned in section 7.3.1, I decided to create subcategories of Accuracy that would analyse Accuracy qualitatively and quantitatively, and which English teachers were familiar with, given that the results of coding using the scheme were to be used as a direct basis for developing the RS1. The subcategories were determined on the basis of preliminary analysis of the scripts through repetitive reading. They were identified as what appeared to be potentially features discriminating between neighbouring bands, in

the preliminary analysis.

When I tried to identify the subcategories I found that some of them caused the clause to be unintelligible, while others did not. As the diaries and think-aloud protocol in Chapter Six revealed that the former type of error elicits a more negative response from raters than the latter, I thought that differentiation needed to be made between two types of error. Therefore, I divided the subcategories into two groups: 'Intelligible' and 'Unintelligible'.

### (1.1) 'Intelligible'

From preliminary analysis of the scripts, it became clear that some of grammatical errors observed in the scripts still did not greatly affect understanding of a clause which include the errors. This category was for such local grammatical errors. I devised the subcategories of this category from the preliminary analysis of the scripts, as shown in Table 7.1 above.

#### (1.1.1) Verbs

#### (1.1.1.2) Distinction between verb types

Cases where any required argument for a verb, adjective or preposition (Haegeman, 1994), such as the complement of an intransitive verb or the object of a transitive verb is omitted, fall into this category.

#### (1.1.2) Tense

Tense in English, which is realised by verb inflection, is a grammatical category to "indicate a particular point in time or period of time" (Collins Cobuild English Grammar, 1990: 245). On the other hand, aspect is a grammatical category that reflects the way in which the meaning of a verb is viewed with respect to time. English has two aspects: the perfect and the progressive. Bardovi-Harlig (2000) explains that "tense locates an event or situation on the time line" whilst aspect "provides a means of expressing one's view of a situation or event" (p.10).

Tense and aspect, therefore, are different grammatical concepts. Nonetheless, aspect is dealt with alongside tense in many textbooks and grammar books for teaching/learning English in Korea, and is considered as a part of tense.

Consequently, for the coding scheme in this study, these two categories were merged into one. Otherwise, I felt that the analysis, and in turn the RS1 that would be developed on the basis of the analysis using the coding scheme, might not have been familiar/accessible to Korean teachers.

Therefore, the following cases were counted in this category: choice of tense inappropriate for a given context, and wrong form of tense (e.g., *seen* instead of *have seen*, *he singing* instead of *he is singing* and *bringed* instead of *brought*).

### (1.1.5) Conjunctions and Relatives

In the preliminary analysis, the observed errors with regard to conjunctions and relatives were found to be two kinds: syntactic errors (e.g., *The place is well known to foreigners, many people visit there*) and semantic errors (e.g., *I would like to introduce the places* ***when*** *I had great times* instead of ***where*** *I had great times*).

### (1.1.10) Auxiliaries

### (1.1.10.2) Other

This category aimed to deal with errors in the use of modal auxiliaries. From the preliminary analysis of the scripts, it was found that errors with modal auxiliaries could be divided into syntactic errors and semantic errors. Syntactic errors are errors that result in ungrammatical clauses including the auxiliary (e.g., *He* ***could went*** *to there during the summer holidays*), whilst semantic errors are the result of a semantically inappropriate choice of auxiliary in a given context (e.g., *You* ***will*** *go to the place during this season to enjoy tinted leaves in the mountain* instead of *You* ***could*** *go to the place during this season to enjoy tinted leaves in the mountain* for a given context).

### (1.1.12) Vocabulary and Phrase

Many rating scales specify that an extensive vocabulary should increase the score in this dimension. Use of extensive vocabularies might mean both the range/size and appropriateness of vocabularies in use. However, the term could be understood only as the former of these two. However, I do not agree that vocabulary size should be in some cases a criterion for assessing writing proficiency. Instead, I believe that a more

important criterion than vocabulary size *per se* and just the use of advanced or sophisticated words is the appropriateness of words in a context. This is especially the case in circumstances where the writing topic is given beforehand, as in this study, or where a dictionary is available during writing. For Vocabulary and Phrases, therefore, I intended to deal with the aspect of appropriateness of vocabulary and phrases.

Additionally, I found that many low-level learners created words or literally translated Korean words into English. These appeared to be worthy of analysis, and potentially one of the distinctive categories between levels of writing proficiency.

As a result, three subcategories were created: Word coinage (e.g., *east big door* which was literally translated into English for a proper noun, *Dongdae-mun* and *existanted* which was invented, looking as if it were an English word); Inappropriate words and phrases (e.g., the use of *appliance* instead of *tool*); and Words that are literally translated from Korean or phrases that are either ungrammatical or literally translated from Korean (e.g., *the thousand year Silla Kingdom* instead of *a thousand years of Shilla Kingdom, our school* instead of *my school*[57] and *here's season* instead of *the season here*).

### (1.1.13) Clauses that are either ungrammatical or literally translated from Korean

The focus of this category was the same as in the previous category, category (1.1.12.3): Words that are literally translated from Korean, or phrases that are either ungrammatical or literally translated from Korean. The difference between this category and the previous one lies in the unit to be analysed. The previous one dealt with the word and phrasal level, whilst this category was concerned with the clause level. The reason for differentiating between these two unit levels was that the degrees of ungrammaticality were perceived differently, with the errors at clause level being more serious than those at word and phrasal level, and I believed that this difference should be recognised for the sake of coding.

---

[57] In Korean language the concept 'our' is commonly used instead of 'my'. For example, Koreans say in Korean *our Mum* instead of *my Mum* and *our home* instead of *my home*. Therefore, writing like this in English could be the result of literal translation from Korean into English.

### (1.2) 'Unintelligible'

All the categories above are grammatical errors that do not seriously affect the meaning of a clause, so that readers still find the clause containing the errors intelligible. On the other hand, the scripts also contained errors that made the clause unintelligible. I classified these types of errors under the heading 'Unintelligible'. From the preliminary analysis, the errors were sorted into three subcategories: Errors in clause construction; Use of unintelligible vocabulary; and Other.

### 7.3.2.2 Fluency

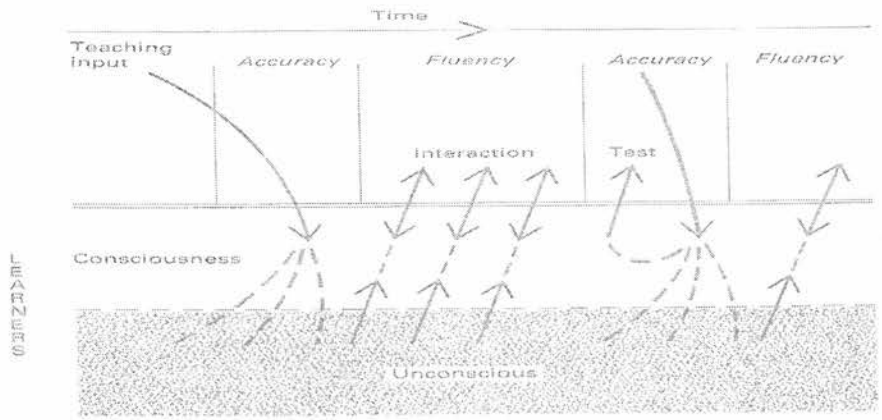
It has been very common to refer to accuracy in the teaching and testing of writing. However, for the last twenty years or more the trend in L2 teaching/learning has moved from grammar-focused teaching to communication-focused teaching. As a result, the focus on accuracy, for which grammar-based methodologies such as Presentation-Practice-Production (P-P-P) and drills were used, has given way to a focus on fluency, which employs activities such as interactive small group work (Richards, 2002). Even so, it has not been common to discuss fluency in the field of the teaching of writing. According to Brumfit (1984), this is due to the characteristics of writing. Whilst production in speaking can be adjusted in response to the apparent incomprehension of the interlocutor, this is not the case with writing. The writing becomes public independently of the writer and is judged by decontextualised criteria. In this context, therefore, teachers and students attend to accuracy, especially at the early stages of writing, contending that learners should first master the language system. As a result, little attention has been paid to fluency in writing assessment. However, this is no longer the case. Studies such as Wolfe-Quintero *et al.* (1998) have introduced the concept of fluency into writing assessment, and it is dealt with in the English Writing course in Korea as well .

Although it is generally recognised that fluency-based pedagogy is task-focused (Richards, 2002),[58] there are various definitions of it. Brumfit (1984) defines it as

[58] Whilst fluency-based pedagogy is task-focused, accuracy-based pedagogy is grammar-focused. Richards sees the characteristics of these two pedagogies as follows:

"natural language use, whether or not it results in native-speaker-like language comprehension or production" (p. 56) and at the same time "the maximally effective operation of the language system so far acquired by the student" (p. 57). His concept of fluency is best shown in comparison with that of accuracy, as in the following diagram, shown here as Figure 7.1 below.

Figure 7.1 Accuracy and Fluency (quoted from Brumfit, 1984: 66)




[Originally Fig. 1d]

According to Leeson's (1975) definition of fluency, with speaking rather than writing in mind, it is "... the ability of the speaker to produce indefinitely many

| |
|---|
| ***Grammar-Focused Activities***<br>Reflect typical classroom use of language<br>Focus on the formation of correct examples of language<br>Produce language for display (i.e., as evidence of learning)<br>Call on explicit knowledge<br>Elicit a careful (monitored) speech style<br>Reflect controlled performance<br>Practice language out of context<br>Practice small samples of language<br>Do not require real authentic communication |
| ***Task-Focused Activities***<br>Reflect natural language use<br>Call on implicit knowledge<br>Elicit a vernacular speech style<br>Reflect automatic performance<br>Require the use of improvising, paragraphing, repair, and reorganisation<br>Produce language that is not always predictable<br>Allow students to select the language they use<br>Require real communication |

(quoted from Richards, 2002: 37)

sentences conforming to the phonological, syntactical and semantic exigencies of a given natural language on the basis of a finite exposure to a finite corpus of that language" (p.136). Towell *et al.* (1996) characterise fluency in terms of automaticity, stating that "fluent second language production requires that knowledge (from whatever source) be proceduralised" (p.85). Lennon (1990) more narrowly defines the term, in terms of the rate and length of output. Based on Lennon's view of fluency, Wolfe-Quintero *et al.* (1998) also define it in terms of rate and length. That is, they contend that:

> [f]luency means that more words and more structures are accessed in a limited time, whereas a lack of fluency means that only a few words or structures are accessed [....]. Fluency is not a measure of how sophisticated or accurate the words or structures are, but a measure of the sheer number of words or structural units a writer is able to include in their writing within a particular period of time (p.14).

Fillmore (1979), however, views fluency more comprehensively than the researchers mentioned above (essentially as to how well people speak their language), and he suggests that the term 'fluency' covers a wide range of language abilities. He proposes four kinds of fluency. First, fluency means "the ability to talk at length with few pauses, the ability to fill time with talk" (p.93). He notes that people who make their living from speaking, such as disc jockeys or sports announcers, are examples of this kind of fluency. Second, fluency means "the ability to talk in coherent, reasoned and "semantically dense" (p. 93) sentences. The main ingredient in this kind of ability appears to be a mastery of the semantic and syntactic resources of the language" (p. 93). He takes William Buckley and Noam Chomsky as examples. Third, fluency means "the ability to have appropriate things to say in a wide range of contexts" (p. 93). He deems that the person with this kind of fluency is at ease in various kinds of conversational settings, whilst the person who is not fluent in this aspect is fluent only in certain familiar settings. Finally, fluency means "the ability [....] to be creative and imaginative in one's language use, to express one's ideas in novel ways, to pun, to make up jokes, to attend to the sound independently of the sense, to vary styles, to create and build on metaphors, and so on" (p. 93). In sum, he defines fluency in terms of, if I express it for the case of second language writing, producing written language rapidly, coherently, appropriately and creatively. After

classifying the term of fluency in these four aspects, he maintains that "[t]he maximally gifted wielder of language, then, is somebody who has all of these abilities" (p. 93).

Once fluency had been defined, a method for coding the features observed in the scripts in terms of fluency needed to be determined. This was not easy. For this, several methods suggested in the previous studies were considered. One uses objective measurements such as a type-token ratio for lexical items and words per T-unit, words per clause and words per error-free T-unit (Wolfe-Quintero *et al.*, 1998). However, all of these methods appear to result in a narrowing down of the concept of fluency. Lennon (1990) also narrowed down the definition of fluency and measured it in terms of rate and length. Therefore, he counted the overall number of words written in a given time.

However, these kinds of operationalisations and the rating scales which would be produced on the basis of the operationalisations were not considered appropriate for this study, as they would make the RS1 look unfamiliar to the English teachers. Therefore, as with Accuracy, I decided to establish the subcategories and quantitatively and qualitatively analyse the corpora. I established four subcategories, reflecting the features observed in the writing corpora and the definition of Fillmore (1979). As a result, four subcategories for Fluency were used: Quantity (length), Coherence, Cohesive devices and Advanced language. Of these four subcategories, Quantity is for the first kind of the four kinds of fluency defined by Fillmore (1979), Coherence and Cohesive devices are for his second kind of fluency and Advanced language is for the third kind. Both 'registers' that could be included with regard to the third and fourth kind of fluency could have been added to, but since the Korean students' scripts were not differentiated from each other in terms of register, as the raters mentioned in their diaries (see section 6.3.5.4.2) and as I found from the preliminary analysis, this did not appear to be useful for the purpose of analysis. It was the same case with the fourth kind of fluency.

### (2.1) Quantity

I included the category Quantity as a subcategory of Fluency on the basis of both preliminary analysis of the scripts and Fillmore's (1979) first kind of fluency, "the

ability to talk at length with few pauses, the ability to fill time with talk". This kind of fluency appeared to be based on spoken language. For written language, this would be the ability to write at length with few hesitations, the ability to fill time with writing. In other studies which investigated fluency in written language, fluency is seen in this manner. Additionally, the obtained scripts also seemed distinguishable in terms of quantity. Therefore, it was considered legitimate to include this aspect in the category of Fluency in this analysis. There might be a concern that length would become the most salient feature and have too much influence on other categories when it was fed into the RS1, just as in the case of the e-rater, as discussed in section 4.6.2, but I did not think this would be the case. Firstly, this was because even though length was included in the coding categories at this stage, it might not be determined as a statistically significant distinctive feature at all after statistical analysis. At the current stage, this category was included among the coding categories as one of the potential distinctive categories, on the basis of preliminary analysis and the literature review. Secondly, this was because even though it did turn out to be a distinctive feature after all, it was supposedly just "one" of two or more distinctive features in "some" cells in the Fluency column in the RS1, and raters would be guided to focus on length when it was suggested as one of the distinctive features in some cells. Therefore, it would not be likely to operate as a surrogate for all other categories, but it was expected that the raters would try to take advantage of the suggested distinctive features rather than only one specific feature such as length. The phenomenon of the dominance of length might apply to holistic scales that include all features in "one" cell for each band and could therefore lead raters to focus on parts of them selectively by themselves, but hopefully, it would not be the case with the RS1, which was to be developed into the form of analytic scales. Length might even dominate with analytic scales if they have various unweighted descriptors for each cell, and consequently tend to make raters pay attention to some of them as mentioned above, but supposedly, this would not be the case with the RS1, which was to be developed to provide the salient and distinctive features of each cell in the scale to focus on. Therefore, the raters' consideration of length would be limited to Fluency rather than all three assessment categories, and furthermore, might only apply to some of the bands (of course, the possibility that it might turn out to be the

case with all of the bands could not be excluded at this stage. This were to be revealed through statistical analysis). In conclusion, it was expected that when length was established as one of the distinctive features for a pair of neighbouring bands for Fluency in the RS1, the raters would focus on it, but if it was not, they would focus instead on other suggested distinctive features.

For consistent analysis of the quantity of the scripts, I established four subcategories to discriminate the scripts by as follows: Less than 33% of about two hundred words (less than sixty-six words);[59] Between 33% and 75% (between sixty-six and one hundred and fifty words); Around 100% (around two hundred words); More than 150% (more than three hundred words).

### (2.3.3) Use of advanced connectors[60]

This category aimed to deal with the clauses headed by subordinators such as *in that, such that, in the event that, assuming that, given that, as far as, according as, in case, as.. so, so... that, less... than, hardly... when, even though, although, as if* and *given that* and transition words such as *subsequently, actually, therefore, thereupon, nevertheless, on the contrary, as a result, in addition, as a result of that,* and *instead of that*. As these were judged to be beyond the level of middle school English in Korea, which covers coordinators and subordinators at beginners level, such as *if, when,* and *as*, I called them advanced connectors.

### (2.4.1) English-like vocabulary, phrases and lexical phrases

In the writing corpora, many low and intermediate students were found to use what is called Konglish in their English writing. Students at advanced levels appeared to command more English-like expressions in terms of vocabulary, phrases *(e.g., I find the book interesting* rather than *I think the book is interesting*, with the former less common in Korean students' English and the latter common) and lexical phrases (e.g.,

---

[59] After the pilot study on writing tasks, the prompts were revised so that they did not include the minimum length requirement for the main study (see section 5.5.1). However, I was told by T1, T2 and T3 that since there was no specification on the minimum length in the prompts for the main study, some of the subjects asked the teachers about it and the teachers publicly told them that two hundred words would be enough, given the requirement in the previous pilot study. Accordingly, even though the minimum quantity was not addressed in the prompts for the main study, I was able to establish two hundred words as a standard average length for the purpose of the analysis.

[60] I generally named coordinators, subordinators and transition words as connectors for this study.

***It seems*** *that the weather is cold today,* ***What I would like to argue is that*** *nature should be preserved,* ***I'm sure*** *you will enjoy traditional Korean food*).

Lexical phrases can be understood as 'formulaic chunks' in Widdowson's (1989) terms, while Nattinger and DeCarrico (1992) define them as " 'chunks' of language of varying length, phrases like *as it were, on the other hand, as X would have us believe*" (p. 1). In addition to this, they explain the term as "multi-word lexical phenomena that exist somewhere between the traditional poles of lexicon and syntax" (p. 1). According to these two authors, the term is used to cover a range from short, relatively fixed frames such as *hi / hello* in an informal letter and *a____ ago* to longer phrases or clauses such as *If I X, then I Y, It has been asserted that X*, and *What I want to recommend is that X.*

I established two subcategories to code this category: Uses them once or twice; and Uses them more than three times. I thought that the demarcation according to frequency could be one of the features between proficiency levels in the preliminary analysis, and the criteria such as "once or twice" and "more than three times" were derived from the preliminary analysis.

### (2.4.2) Good clause construction and good expansion of clauses through fluent use of adjective/adverbial clauses

This category dealt with the clause constructions most appropriate for conveying the intended meaning, following information structure rules[61] and extending sentences

---

[61] A prepositional meaning can be variously expressed. For example, for a meaning of 'A waiter brought them cocktails', there are the following alternatives: 1) A waiter brought them cocktails; 2) A WAITER brought them cocktails; 3) They were brought cocktails by a waiter; 4) Cocktails they were brought by a waiter; 5) It was cocktails that a waiter brought them; 6) What the waiter brought them was cocktails; 7) Bring them cocktails, the waiter did; and 8) There was a waiter who brought them cocktails. Of course there are further alternatives. Of these, in the underlined section of the following text, only 3) and 8) are reasonable alternatives:

> They arrived at the hotel and sat on the terrace. The sun was hot but the canopies gave a pleasant shade. There was a glimpse of the sea through the palm-trees. _____.

On the other hand, the indiscriminate use of alternative forms results in incoherence. The following text shows this point:

> At the hotel they arrived and on the terrace they sat. It was the sun that was hot, but what the canopies gave was a pleasant shade. A glimpse of the sea was through the palm-trees. Brought them cocktails, the waiter did.

> Both of these two examples above can be explained in terms of the information structure rule. In

smoothly, taking advantage of adverbial clauses or adjective clauses.

I established two subcategories for coding a script for this category: Uses them once or twice; and Uses them across the script. The rationale and procedure for making demarcation according to such frequency were the same as in category (2.4.1).

### (2.4.3) Advanced grammar

From the preliminary analysis of the scripts, I drew four main subcategories for advanced grammar: Use of aspect; Use of relative adverbs; Use of *that* clause for complement / subject; and Other.

#### (2.4.3.1) Use of aspect

From the preliminary analysis of the scripts, it was found that many writers mainly used simple tense forms, and that only a small number of relatively high-level scripts included aspects such as present perfect, past perfect, present progressive and present perfect progressive. It seems that many Korean students do not often use aspect, although it is taught at a relatively early stage of English learning in Korea.

It appears from studies on the acquisition of aspect that this phenomenon is not limited to Korean students. As Bardovi-Harlig (2000) notes in the introductory chapter of her book on tense and aspect acquisition for L2 learners, the pedagogy of tense and aspect has received a great deal of attention, to the extent that a large number of language teaching programmes have included mastery of certain tense-aspect forms in their criteria for advancement from one course to another. As many studies on tense-aspect acquisition show (e.g., Bardovi-Harlig, 2000; Lee, 1997), however, it appears that the mastery of tense-aspect forms needs to go through stages, and that the development is slow and gradual. As a result, learners tend to rely on suffixed inflections rather than discontinuous marking such as aux + V –tense inflection (Bhardwaj *et al.*, 1988) for temporal expression.

Consequently, verbal morphologies for tense-aspect forms are not acquired at

---

the first example, since 3) and 8) follow the rule they look reasonable in the text. In the second example, the text is incoherent because it does not follow the rule (Downing & Locke, 1992). Therefore, the information structure rule appears to be one of the factors affecting the conveying of a meaning.

the same time but emerge in a certain order, as Bardovi-Harlig's (2000) study reveals. Among simple past, past progressive, present perfect and pluperfect, the rate of appropriate use of the simple past is higher, followed by past progressive and present perfect, while the last to be acquired is pluperfect. The present perfect is acquired later than the simple past in L2 as well as L1. Bardovi-Harlig summarises the possible explanations suggested for this in the literature under three headings: cognitive development (however, this is not the case with adult L2 learners, unlike L1 learners, according to von Stutterheim and Klein (1987)); morphosyntactic complexity (Gathercole, 1986; Johnson, 1985; Smith, 1980); multiple factors such as semantic complexity and frequency of input.

Given that the tense-aspect system appears to be such a difficult part of language to master for Korean students as well as other L2 learners, it could be considered to play a role in distinguishing between levels of writing proficiency. In addition, it was the case, from the preliminary analysis, that the use of present perfect was generally limited to high levels of scripts. Consequently, this category came to be included in the coding scheme in order to investigate whether there were any cases of correct use of aspect in a script (especially present perfect and past perfect).

#### (2.4.3.2) Use of relative adverbs

From the preliminary analysis, the clauses headed by relative adverbs such as *when, where, why* and *how* were revealed as one of the most common advanced grammatical features in the high-level scripts. Relative pronouns such as *who, which, whom* and *that* were excluded here. Since many writers for this study, even those in low bands, showed cases of grammatical use of relative pronouns,[62] they did not appear to be a feature of advanced level nor a discriminating between writing levels. Therefore, they were excluded here.

### (2.4.4) Multi-word verb phrases

Multi-word verb phrases such as *take up* and *pull off* are frequent in native English

[62] That even low-level students could use relative sentences may be because they attended a foreign language high school. This would not have been the case if they had been students at a general academic high school, whose level of English is lower than that of foreign language high school students.

speakers' production, and may be one of the items that makes a script look English-like. Nonetheless, there are not many of them in Korean students' scripts. Only a very small number of students used these verb phrases. Therefore, as this feature might be one of the features discriminating between levels of writing proficiency, and since it is one that helps make a script look English-like, I included it in the coding scheme.

#### 7.3.2.3 Organisation

This category dealt with the overall structure and the flow of content in a script. Therefore, on the basis of the preliminary analysis of the scripts, this concerned format, overall development, overall content and topic address. With regard to these, I made four subcategories of this category: Paragraphing; Genre format and development; Topic address; and Content.

#### (3.2) Genre format and development

Many kinds of writing, including letters and formal essays, generally consist of three parts: opening, body and closing. For a script to look well organised, it should have these three parts, each of which should be sufficiently developed. Also, the distinctions between these parts need to be clear. To this end, each part needs to follow the appropriate format for the genre of writing, using appropriate lexical phrases. For example, two genres in this study – an informal letter and a formal essay in English – take lexical phrases for their opening part, as shown in Table 7.2 according to Nattinger & DeCarrico (1992) below.

Table 7.2 Lexical phrases for the opening and closing parts for an informal letter and a formal essay (partly quoted from Nattinger & DeCarrico, 1992: 169)

| | **Informal Letter** | **Formal Essay** |
|---|---|---|
| For opening | *hi / hello /dear _____*<br>*do you remember X?*<br>*let me tell you about X* | *for a long time X*<br>*it has been the case that X*<br>*one of the most important controversial _____ (in the recent literature) is X* |
| For closing | *see you later*<br>*well, that's about it (for now)* | *in sum / conclusion*<br>*to summarise* |

[Originally No Table No.]

This category that covers the genre format and development is different from

Paragraphing, which is not concerned with whether a script conforms to the genre format and whether it consists of three parts, but with whether it is paragraphed and whether the paragraphing is right or not.

From the preliminary analysis of the writing data, the various levels seemed distinguishable in genre format and development. It was found that the lower bands of scripts tended not to be well formatted or sufficiently developed, so I included these as one category for Organisation.

To summarise, each coding category was established on the basis of the definition of writing ability implicit in the English Writing course, the theoretical background and the preliminary analysis of the obtained scripts. From the preliminary analysis, it seemed that the categories were potential features discriminating between levels of writing proficiency at this stage (although some of them turned out not to be after statistical analysis of them; see section 7.5). In the next section, I will introduce the coding procedure using this coding scheme.

## 7.4 Coding procedure

For the coding of categories in Accuracy, I counted the errors whose correction would make the clause grammatical (rather than excellent or more advanced), and coded all the observed errors in each script. Errors in clauses can be corrected in various ways, but I wanted to find errors whose correction would make the clause grammatical, given that the Accuracy category is for grammaticality rather than sophistication of usage.

For the coding of categories in Fluency, various coding methods were employed. For a category marked 'C' (e.g., category 2.2.2), I counted errors, as in Accuracy. Of the categories marked 'T', for a category of optional binary choice (e.g., category 2.3.2), when a script satisfied the specification of the category, it was ticked for the category (this coding was converted into '0' for non-ticking or '1' for ticking, for statistical analysis) whilst for categories such as category 2.4.1, only if a script met one of the choices of the category, it was ticked for one of the given choices, i.e., 2.4.1.1 or 2.4.1.2 (this coding was converted for '0' for non-ticking, '1' or '2' for ticking, following the choice numbers, that is, '1' when category 2.4.1.1 was chosen

and '2' when category 2.4.1.2 was chosen, for statistical analysis). On the other hand, for a category of obligatory multiple-choice (e.g., category 2.1), every script was necessarily ticked for one of the choices (this coding was converted into '1', '2' and so on, following the choice number, that is, '1' when category 2.1.1 was chosen, '2' when category 2.1.2 was chosen, for statistical analysis).

For the coding of categories in Organisation, for a category marked 'C', i.e., category 3.2.2, the category occurrence was counted (the numbers were inserted for statistical analysis). For the other categories than category 3.2.2, coding of multiple-choice type was applied: every script was necessarily ticked for one of the choices for each category (this coding was converted into '1', '2' and so on, following the choice number, for statistical analysis).

As can be seen from the coding categories and their coding methods, some appear to give credit, such as "Use of advanced connectors", whilst others appear to downgrade for errors, such as grammatical features in Accuracy. These two seemingly contrasting approaches for coding were adopted in order to reflect the different features or aspects of writing. A script should be downgraded if certain features or aspects of writing are found lacking, such as various grammatical features in Accuracy, because they are needed to make clauses comprehensible and intelligible. However, the lack of other features in a script, such as the use of advanced connectors, would not make clauses incomprehensible or unintelligible because even basic connectors can convey the meaning, even if not finely. Therefore, the lack of these features should not cause a script to be downgraded, although their presence could attract credit. This is why these two different – positive and negative – approaches for coding were adopted and fed into the descriptors of the RS1. This aspect did not appear to be problematic or confusing. When I coded a script using the coding scheme, since I just needed to tick whenever appropriate[63] for each coding category, by investigating whether the description of the category appears to apply to the script, rather than, for example, indicating '+' for giving credit or '-' for downgrading, I could easily maintain consistency with these two approaches. Furthermore, as can be seen from the feedback on both the RS1 and the RS2 in

[63] After ticking whenever necessary (e.g. in the category Agreement whenever errors for this feature were found in a script), the ticks were added up to give a score, such as '4' for SPSS statistical analysis.

sections 7.6.3, 8.3 and 8.4, none of the raters reported that they were confused by the fact that the RS2 included both positive and negative wording. Furthermore, since positive and negative elements in descriptors in a scale can be observed in other rating scales, apart from 'can do' scales such as the Common European Framework (Council of Europe, 2001) and Association of Language Testers in Europe (ALTE) Framework (Council of Europe, 2001), which have only positive descriptors, it does not appear to be peculiar to the RS1 and RS2.

For the coding of these three main categories, I devised a coding sheet including all the coding categories, which is shown in Appendix 8. One sheet was used per script.

As mentioned in section 5.6.2, I investigated the quality of data coding after developing the coding categories and the coding sheet, and before starting coding of the scripts. After this investigation, I coded the whole scripts. An example of a coded script can be seen in Appendix 9.

After the coding of the writing samples, all the coding was statistically analysed to draw out the distinctive features between bands that would be the basis for developing the rating scale. In the next section, I will discuss the statistical analysis in detail.

## 7.5 Procedure for the statistical analysis

Through this statistical analysis, I tried to find features discriminating between neighbouring bands (i.e., Bands 2 vs. 3, Bands 3 vs. 4, and so on). First, after crosstables for all variables were created, variables that were not statistically significant were discarded and statistically significant variables were determined by Test of independence, that is chi-square.[64] P-value was considered to be at the level of 0.10 rather than 0.05 or less in order to make it less conservative, and to avoid excluding otherwise useful distinctive variables and ending up having few distinctive variables to feed into the RS1.

As a result, fifteen variables in Accuracy, sixteen variables in Fluency and nine variables in Organisation were chosen as statistically significant variables. I then

[64] The sample for Band 1 was only one of the three hundred and ninety samples. Since only one sample was not representative for the band, the statistical analysis to find the distinctive features was done for Bands 2 to 6.

compared the modes of each band for each significant variable, and when the mode for a band was significantly different from that for its neighbouring band, I allocated the variable as a distinctive variable for the two bands. For example, for variable 2.2.1 (Disconnected and incoherent sentences for more than 50% of the whole script), this variable was determined as significant because its p-value (.0001) was smaller than .10 (as discussed above, p-value in this analysis was considered to be at the level of .10). Therefore, this variable was found to be a significantly distinctive feature for some pairs of neighbouring bands. To decide which pairs of neighbouring bands were significantly distinguished by this variable, I consulted frequency, i.e., modes of each band for this variable in the Crosstabulation. As can be seen in Table 7.3 below, the mode for Band 2, i.e., '1' is significantly distinct from that of Band 3, i.e., '0', whilst it is not the case with the other bands. Therefore, this variable was allocated as a distinctive variable between Bands 2 and 3, as can be seen in Table 7.5 in section 7.5.1.

Table 7.3 Crosstable for variable 2.2.1

| Band | Disconnected and incoherent sentences for more than 50% of the whole script | | Total |
|---|---|---|---|
| | No (0) | Yes (1) | |
| 2 | 12 (30%) | 28 (70%) | 40 (100%) |
| 3 | 101 (82%) | 22 (18%) | 123 (100%) |
| 4 | 112 (94%) | 7 (6%) | 119 (100%) |
| 5 | 87 (98%) | 2 (2%) | 89 (100%) |
| 6 | 18 (100%) | 0 (0%) | 18 (100%) |
| Total | 330 (85%) | 60 (15%) | 390 (100%) |

NB The modes of each band for this variable are '1' for Band 2 and '0' for Bands 3,4,5 and 6 in the table.

For another example, as for variable 2.1 (Quantity), this variable was determined as significant because its p-value (.0001) was smaller than .10. Therefore, this variable was found to be a significantly distinctive feature for some pairs of neighbouring bands. To decide which pair of neighbouring bands this variable significantly distinguished, I consulted frequency, i.e., modes of each band for this variable in the Crosstabulation. As can be seen in Table 7.4 below, the modes for Bands 2, 3, 4, 5 and 6 are 2, 3, 4, 4 and 4 respectively, as can be seen in Table 7.4.

Therefore, this variable discriminates between Bands 2 and 3, as can be seen in Table 7.5 in section 7.5.1, and between Bands 3 and 4, as can be observed in Table 7.6 in section 7.5.2. Thus, this variable was allocated as one of variables discriminating both between Bands 2 and 3 and between Bands 3 and 4. In this manner, I established a set of distinctive variables for each pair of neighbouring bands, using all of the selected statistically significant variables.

In the next sections I will discuss the distinctive variables established in this manner, according to each pair of neighbouring bands.

Table 7.4 Crosstable for variable 2.1

| Band | Quantity | | | | Total |
|---|---|---|---|---|---|
| | Less than 33% of minimum quantity of around 200 words (1) | Between 33% and 75% of minimum quantity of around 200 words (2) | Around 100% of minimum quantity of around 200 words (3) | More than 150% of minimum quantity of around 200 words (4) | |
| 2 | 3 (7.5%) | 28 (70%) | 8 (20%) | 1 (2.5%) | 40 (100%) |
| 3 | 0 (0%) | 34 (27.64%) | 48 (39.02%) | 41 (33.33%) | 123 (100%) |
| 4 | 0 (0%) | 12 (10.08%) | 31 (26.05%) | 76 (63.87%) | 119 (100%) |
| 5 | 0 (0%) | 0 (0%) | 25 (28.09%) | 64 (71.91%) | 89 (100%) |
| 6 | 0 (0%) | 0 (0%) | 4 (22.22%) | 14 (77.78%) | 18 (100%) |
| Total | 3 (0%) | 74 (19%) | 116 (30%) | 197 (51%) | 390 (100%) |

NB The modes of each band for this variable are '2' for Band 2, '3' for Band 3, and '4' for Bands 4, 5 and 6 in the table.

### 7.5.1 Band 2 vs. Band 3

The distinctive variables between these two bands are as follows:

Table 7.5 Variables discriminating between Bands 2 and 3

| Distinctive Variables (14) | | | | Band 2 | Band 3 |
|---|---|---|---|---|---|
| **Accuracy** | (1.1) Intelligible | (1.1.3)Indication of quantity in nouns and Articles | | 1 | 2 |
| | | (1.1.4)Agreement | | 1 | 2 |
| | | (1.1.5)Conjunctions and Relatives | (1.1.5.2)Semantic error | 0 | 1 |
| | | (1.1.14)Other grammatical errors | (1.1.14.2)Word order in a phrase | 0 | 1 |
| | (1.2)Unintelligible | (1.2.1)Errors in sentence construction | (1.2.1.1)Syntactic error | 0 | 1 |
| **Fluency** | (2.1)Quantity | | | 2 | 3 |
| | (2.2)Coherence | (2.2.1)Disconnected and incoherent sentences for more than 50% of the whole script | | 1 | 0 |
| | (2.3) Cohesive devices | (2.3.4)Errors in the use of number and person of pronouns | | 0 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| | (2.4)Advanced language | (2.4.4)Multi-word verb phrases | | 0 | 1 |
| **Organisation** | (3.1)Paragraphing | | | 1 | 2 |
| | (3.2)Genre format and development | (3.2.1)Opening | (3.2.1.1)Blurred distinction between opening and body or lack of opening | 1 | 0 |
| | | | (3.2.1.3)Development | 0 | 1 |
| | | (3.2.2)Body | (3.2.2.1)Number of points | 1 | 3 |
| | (3.4)Content | | | 3 | 4 |

NB. The figures (e.g., 0,1,2,3) for each band indicate the modes of occurrences of each variable. Thus, for the variable marked 'C' they mean the mode of occurrences, whilst for the variable marked 'T' they mean the choice number that was most frequently chosen.

As can be seen in the table above, in terms of Accuracy, Band 2 scripts had fewer intelligible grammatical errors and unintelligible grammatical errors than those in Band 3. However, this could be because the average length of Band 2 scripts was shorter than that of Band 3 scripts.

For Fluency, in terms of length, Band 2 scripts are generally between 33% and 75% of two hundred words, whilst those of Band 3 are around 100% of two hundred words. The main distinctive feature between the bands in Fluency was related to Coherence. Most of the Band 2 scripts were incoherent and disconnected, to the extent that over 50% of these scripts were disconnected sentences, whilst the scripts of Band 3 were not. Additionally, with regard to the use of advanced language, unlike the writers of Band 2, the writers of Band 3 used multi-word verb phrases.

For Organisation, most of the Band 2 scripts were not paragraphed, their opening part was not distinguished from their body, and the body generally contained only one point. On the other hand, most of the Band 3 scripts were paragraphed, even though there were some errors with paragraphing, and the opening and body of the scripts were distinguished, even though they were not sufficiently developed. With regard to Content, most of the Band 2 scripts included very little of the required content, with only one point, whilst those of Band 3 on the whole included the required content, with an average of three points.

### 7.5.2 Band 3 vs. Band 4

Table 7.6 below shows the distinctive variables between Bands 3 and 4.

Table 7.6 Variables discriminating between Bands 3 and 4

| Distinctive Variables (5) | | | | Band 3 | Band 4 |
|---|---|---|---|---|---|
| **Accuracy** | (1.2)Unintelligible | (1.2.1)Errors in sentence construction | (1.2.1.1)Syntactic error | 1 | 0 |
| **Fluency** | (2.1)Quantity | | | 3 | 4 |
| **Organisation** | (3.2)Genre format and development | (3.2.1)Opening | (3.2.1.3)Development (quantity) | 1 | 3 |
| | | (3.2.3)Closing | (3.2.3.3)Development (quantity) | 1 | 2 |
| | (3.3)Topic address | | | 1 | 5 |

For Accuracy, the presence of syntactic errors in sentence construction that led to unintelligible clauses was likely to make the difference between scripts being put into Band 3 or Band 4. Most of the Band 3 scripts included these kinds of errors, whilst on the whole those of Band 4 did not.

For Fluency, the length of scripts tended to differentiate Band 3 from Band 4. Most Band 3 scripts were usually around the full two hundred words (100%), while Band 4 scripts were more than 150% of two hundred words.

Pertaining to Organisation, Band 3 scripts were likely to be insufficiently developed in terms of both opening and closing, to the extent that each part was less than two sentences long. Band 4 scripts were slightly longer and more developed than those of Band 3, with the opening more than three sentences long, and the closing between two and three sentences. With regard to topic / purpose address in the opening stage, it is also worth noting that many Band 3 scripts did not explicitly address the topic, whilst in those of Band 4 the topic / purpose of the discourse was appropriately addressed in the opening part.

### 7.5.3 Band 4 vs. Band 5

From the analysis, the variables in Table 7.7 below were revealed to be distinctive between Bands 4 and 5.

Table 7.7 Variables discriminating between Bands 4 and 5

| Distinctive Variables (14) | | | | | Band 4 | Band 5 |
|---|---|---|---|---|---|---|
| **Accuracy** | (1.1)Intelligible | (1.1.1) Verb | (1.1.1.1)Distinction between finite and non-finite verbs/omission or repetition of finite verbs | | 1 | 0 |
| | | (1.1.5)Conjunctions and Relatives | (1.1.5.2)Semantic error | | 1 | 0 |
| | | (1.1.7)Voice and Participles | | | 1 | 0 |
| | | (1.1.9)*To*- or bare infinitives and Gerund | | | 0 | 1 |
| | | (1.1.10)Auxiliaries | (1.1.10.2)Other | (1.1.10.2.1)Syntactic error | 1 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Fluency** | (2.2)Coherence | (2.2.2)Local lack of coherence | (2.2.2.1.2)Due to its irrelevance to the previous sentence | | 2 | 1 |
| | (2.3)Cohesive devices | (2.3.1)Little repetitions using substitution words | | | 0 | 1 |
| | | (2.3.3)Use of advanced connectors | | | 0 | 5 |
| | | (2.3.4)Errors in the use of number and person of pronouns | | | 1 | 0 |
| | (2.4)Advanced language | (2.4.3)Advanced grammar | | (2.4.3.4)Other | 0 | 1 |
| **Organisation** | (3.1)Paragraphing | | | | 2 | 3 |
| | (3.2) Genre format and development | (3.2.2)Body | (3.2.2.2)Number of insufficiently developed points | | 2 | 1 |
| | | (3.2.3)Closing | (3.2.3.1)No closing | | 1 | 0 |
| | | | (3.2.3.3)Development (quantity) | | 2 | 4 |

Pertaining to Accuracy, writers in both Bands 4 and 5 were unlikely to make grammatical errors that rendered a clause unintelligible. Rather, they were likely to make local errors which did not affect the intelligibility of clauses. However, they were found to be different from each other in terms of the quantity of errors, with Band 4 scripts likely to have more errors than those of Band 5. The errors spanned various categories, i.e., verb, conjunction, voice, infinitives and auxiliaries, rather than being clustered in any specific categories.

With regard to Fluency, scripts in both bands showed a tendency towards a local lack of coherence due to use of a clause that was irrelevant to the previous clause, although there was a slight difference in quantity between bands. However, they were different from each other in terms of Cohesive devices and Advanced language. Writers in Band 5 were likely to use substitution and connectors beyond the middle school level in Korea, while those in Band 4 were not, and although they used substitution, this contained errors in terms of number and personal pronouns.

With regard to Organisation, Band 4 scripts were likely to have errors in paragraphing whilst those of Band 5 were not. Furthermore, Band 4 scripts were likely to have more points in the body that were not fully developed than those of Band 5. In addition, many of the Band 4 scripts did not have a clear closing part. Even though there was a closing part, it was likely to be insufficiently developed in many Band 4 scripts, to the extent of being less than three sentences long. The closing part of Band 5 scripts was more likely to be reasonably developed and rounded off.

### 7.5.4 Band 5 vs. Band 6

The distinctive variables between Bands 5 and 6 revealed in the analysis are

represented in Table 7.8 below.

Table 7.8 Variables discriminating between Bands 5 and 6

| Distinctive Variables (21) | | | | | Band 5 | Band 6 |
|---|---|---|---|---|---|---|
| **Accuracy** | (1.1)Intelligible | (1.1.3) Indication of quantity in nouns and Articles | | | 2 | 1 |
| | | (1.1.4)Agreement | | | 2 | 1 |
| | | (1.1.5)Conjunctions and Relatives | | (1.1.5.1)Syntactic error | 1 | 0 |
| | | (1.1.6)Distinctions between word classes | | (1.1.6.1)*because/there/for example* (C) | 1 | 0 |
| | | (1.1.11)Spelling, Capitalisation and Punctuation | | (1.1.11.3)Other | 1 | 0 |
| | | (1.1.12)Vocabulary and Phrase | | (1.1.12.3)Words that are literally translated from Korean, or phrases that are either ungrammatical or literally translated from Korean | 1 | 0 |
| | | (1.1.13)Clauses that are either ungrammatical or literally translated from Korean | | | 1 | 0 |
| | | (1.1.14)Other grammatical errors | | (1.1.14.2)Word order in a phrase | 1 | 0 |
| | (1.2)Unintelligible | (1.2.2)Use of unintelligible vocabulary | | | 1 | 0 |
| **Fluency** | (2.2)Coherence | (2.2.2)Local lack of coherence | (2.2.2.1) For an individual sentence | (2.2.2.1.1)Due to insufficient language command | 2 | 1 |
| | | | | (2.2.2.1.2)Due to its irrelevance to the previous sentence | 1 | 0 |
| | | | | (2.2.2.1.3)Due to its unintelligibility | 1 | 0 |
| | | | (2.2.2.2)For more than two consecutive sentences | (2.2.2.2.2)Due to their irrelevance to the previous sentence | 1 | 0 |
| | (2.3)Cohesive devices | (2.3.2)Smooth connection between sentences using cohesive devices well | | | 0 | 1 |
| | (2.4)Advanced language | (2.4.1)English-like vocabulary, phrase and lexical phrases | | (2.4.1.2)Uses them more than three times | 0 | 2 |
| | | (2.4.2)Good clause construction and good expansion of clauses through fluent use of adjective/adverbial clauses | | (2.4.2.2)Uses them across the script | 0 | 2 |
| | | (2.4.3)Advanced grammar | | (2.4.3.1)Use of complex aspect | 0 | 2 |
| | | | | (2.4.3.2)Use of relative adverbs | 0 | 1 |
| | | (2.4.4)Multi-word verb phrases | | | 1 | 2 |
| **Organisation** | (3.2)Genre format and development | (3.2.2)Body | (3.2.2.1)Number of points | | 3 | 5 |
| | | | (3.2.2.2)Number of insufficiently developed points | | 1 | 0 |

As can be seen in the table above, there were fewer grammatical errors in the Band 6 samples than in those of Band 5, which had some errors across various grammatical items and sometimes even in the choice of vocabulary, resulting in an unintelligible clause. This tells us that the scripts in Band 6 are fairly accurate, even

though they have a very small number of local errors.

For Fluency, writers in Band 5 sometimes lacked coherence in both individual sentences and more than two consecutive sentences. In addition, the Band 6 samples were better than those of Band 5 in terms of Cohesive devices and Advanced language. Unlike those in Band 5, Band 6 writers were likely to use cohesive devices such as substitution, and command advanced language such as English-like vocabulary, phrases and lexical phrases more than three times. They often chose very appropriate clause constructions for the intended meaning, used advanced grammar such as the use of aspect (i.e., present perfect and past perfect) and relative adverbs, and used more multi-word verb phrases than writers in Band 5.

For Organisation, the samples in Bands 5 and 6 were similar to each other in many ways. However, whilst those in Band 5 were likely to have at least one point in the body of the script that was not fully developed, those of Band 6 were not.

As discussed above, the variables discriminating between neighbouring bands were derived, and on the basis of both these variables and the characteristics for other variables within the coding scheme, I developed the RS1. I will discuss the scale and the application procedure employed by the three raters in the next section.

## 7.6 The development and application of the RS1

### 7.6.1 The RS1

According to the requirement explained at the very end of section 6.4, the RS1 took the form of an analytic rating scale with three main assessment categories: Accuracy, Fluency and Organisation. It was based on both the construct of writing ability implicit in the English Writing course in Korea and the coding scheme. The scale also had 1 to 6 bands, with Band 6 as the highest band. According to Henning (1996), if there are too many levels in a rating scale, users of the scale are likely to have difficulty in differentiating between bands. On the other hand, if there are too few bands, the scale is not likely to differentiate between the proficiency levels of test-takers. Therefore, I believed five or six bands to be reasonable. I chose to make it a six-point-band rather than a five-point one to avoid having a mid-point in the scale. If there is a mid-point in a scale (i.e., Band 3 in a five-point band), users are apt to use

it for scripts around an intermediate level. To give better differentiation between the levels of proficiency, they should be led to choose either above or below the mid-point. Therefore, the RS1 was made to have six bands in it.

I needed to determine some rationale when developing descriptors of the scale. First, I tried to include all of the obtained statistical results, i.e., the features discriminating between neighbouring bands. These features were intended to help raters grasp the difference between neighbouring bands.

Second, I tried to reflect not only the features that discriminate between bands, but also the characteristics of some of the other variables (that seemed necessary) within the coding scheme revealed by the statistical analysis. For example, for Band 2 Fluency, I included not only "Global lack of coherence across over 50% of sentences in the whole script. This writing is merely the enumeration of incoherent and disconnected sentences" which was the feature distinguishing itself from Band 3, but also "Few clauses show either good choice of clause construction or clause expansion. No good expression. Very plain level of English. Repetition of words without using pronouns and, if any, errors in their use", none of which actually distinguish Band 2 from its neighbouring band, but which reflect some of the findings on the basis of modes for other variables from Crosstabulations. By doing this, I intended to help raters grasp the overall picture of each band and thus the entire scale, as well as features distinguishing from its neighbouring bands.

Third, in addition to these discriminating or general features for other variables, each cell was designed to have supplementary aids to help clarify certain descriptors that seemed to require further explanation. For descriptors that include terms which are either specific to this scale or not self-explanatory, I added some brief guidelines in bullet points at the bottom of the features discriminating between bands in each cell of the scale.

Finally, to indicate quantitative differences between the bands that the statistical results show, I used quantifiers and frequency adverbs. I used words such as "few", "a few", "some" and "many" stated in the intended order of frequency, choosing the most appropriate words for each band on the basis of the statistics obtained through statistical analysis and my empirical repetitive reading of the scripts assigned to each band.

Following the rationale outlined above and reflecting on the findings from the statistical analysis discussed in previous sections in this chapter, I constructed the RS1 as shown in Table 7.9 below.

Table 7.9 The RS1

| | Accuracy | Fluency | Organisation |
|---|---|---|---|
| Band 6 | **Few grammatical errors or a few, if any, local grammatical errors. No phrase or clause that is either unintelligible or has problems in construction. Or more accurate than this.**<br><br>●Local grammatical errors means errors in, for example, number, articles and agreement. | **More than minimum length. Lack of coherence on less than a few occasions due to insufficient language command. No case that is incoherent due to irrelevant content. Generally good coherence. Outstanding cohesion between sentences. Smooth flow between sentences due to good clause construction and expansion for rich expression. Advanced level of language using advanced grammatical features. Or more fluent than this.**<br>●Cohesion between sentences means good use of pronouns and connectors.<br>●Good clause expansion means clause expansion using adjective/adverbial phrases/clauses for rich expression and good communication.<br>● Advanced grammatical features mean perfect aspect, relative adverbs, clause construction beyond middle school level; whilst advanced level of language means using multi-word verbs, connectors beyond middle school level, and English-like vocabulary/phase and clause construction/ lexical phases. | **Good organisation and paragraphing. Clear topic address. Good development in all points of opening, body, and closing. Covers the required content well. Or better organised than this.**<br><br>●Good development in all points of opening, body, and closing means that each point is quantitatively well developed to 3 sentences or 2 and a half lines (note that even if a point is longer than this, if it is not rounded off in terms of content, it needs to be considered as insufficiently developed). |
| Band 5 | **Some errors across various kinds of grammatical features. A few phrases and clauses which have problems in construction. Few clauses which are unintelligible due to wrong choice of vocabulary. However, no clause which has errors in verbs.**<br><br>●Errors in verbs include omission of finite verb of a clause, errors in distinction between finite and nonfinite verbs, and errors in voice and participles. | **More than minimum length. A few local lack of coherence. No excellent smooth flow between sentences due to failure to use appropriate clause construction. Neither good clause expansion nor good expression. However, good cohesion between sentences due to correct use of pronouns, little repetitions and use of connectors beyond middle school level.**<br>●Little use of English-like vocabulary or phrase<br>●Little use of perfect aspect and relative adverbs | **Covers the required content well. Clear topic address. Good development in opening and closing. But one (or two) point(s) in the body of the script insufficiently developed. However, good organisation on the whole. Few errors in paragraphing.** |
| Band 4 | **Many errors across various kinds of grammatical features, especially in verbs. Some clauses which are unintelligible due to wrong choice of vocabulary. However, no clause which is unintelligible due to serious errors in clause construction.** | **Some local lack of coherence due to both insufficient language command and irrelevance to the previous sentence. Plain level of language which does not use good clause expansion, good expression and appropriate clause construction. Repetition of words without using** | **Covers the required content well. Some of the points in the body of the script are insufficiently developed. Errors in paragraphing. Errors in overall organisation due to unclear closing part. However, well-developed opening and clear topic address.** |

| | | | |
|---|---|---|---|
| | | **pronouns and, if any, errors in their use. <u>However, able to write more than the minimum length.</u>** | ●Unclear closing part means it is either omitted or underdeveloped. |
| Band 3 | **Many errors across grammatical features. Some clauses which are unintelligible due to either wrong choice of vocabulary or serious errors in clause construction. <u>However, only some clauses are unintelligible.</u>** | **Around required length. Few clauses which show either good choice of clause construction or clause expansion. No good expression. Very plain level of English. Repetition of words without using pronouns and, if any, errors in their use. Frequent local lack of coherence, <u>but not a general lack at the level of overall script</u>.** | **Errors in paragraphing. Errors in overall organisation due to unclear opening and closing parts. Despite there being many points in the body of the script, most of them are insufficiently developed. Omission of topic address. <u>However, covers the required content well.</u>** |
| Band 2 | **A large number of errors across all kinds of grammatical features including verbs with almost all of the clauses. Many phrases or clauses which have problems in construction. Many clauses which are unintelligible due to either wrong choice of vocabulary or serious errors in clause construction.**<br>●You do not have to assess Accuracy in cases where both Fluency and Organisation of the script in question are assessed at Band 2.<br>●The script of this band tends to follow English word order, i.e. S+V+O/C rather than Korean word order, i.e. S+O/C+V even though it is very poor with a great many grammatical errors. | **Less than minimum length. Few clauses show either good choice of clause construction or clause expansion. No good expression. Very plain level of English. Repetition of words without using pronouns and, if any, errors in their use. Global lack of coherence across over 50% of sentences in the whole script. This writing is merely the enumeration of incoherent and disconnected sentences.**<br>●Less than required length means around 33~70% of required length. | **No paragraphing and overall very poor organisation. Opening and closing parts either lacking or unclear. If any, very insufficiently developed. Extremely insufficient number of points in the body of the script. Extremely insufficient content.**<br>●Extremely insufficient content means that although the writing in this band addresses the required content, it is not communicated clearly, either due to poor development of each point or due to lack of overall coherence.<br>●Extremely insufficient development of opening, body and closing means that it is less than 2 sentences/one and a half lines quantitatively. |
| Band 1 | **Less accurate than Band 2** | **Less fluent than Band 2** | **More poorly organised than Band 2** |

NB. Underlined parts in each cell are the features that distinguish this band from the one below, whilst non-underlined parts are either the overall features of the band or features that distinguish this band from the one above in terms of each assessment category. The bullet points in the cells are supplementary explanations of each cell.

I prepared the RS1 for use by the three raters. There were two purposes for this: to get feedback from them and revise the rating scale on the basis of the feedback; and, given Kenyon's (1997) study (see 4.6.3) and the practical constraints involved, to take advantage of this stage as self-training, a kind of rater training, getting the raters used to the terms and style of the scale. I will discuss this application process in more detail in the next section.

### 7.6.2 Application of the RS1

I used Group H for the application of the RS1 (see section 5.6.3 for more details on

Group H). When I chose the scripts, I intended to select various levels of script according to my subjective holistic scoring,[65] in order to have the raters apply it to various levels of scripts. Additionally, I meant to have the raters apply it to different genres of script in order to see whether there is any difference in the use of the scale depending on genre. Therefore, half of the twenty scripts were on Task 5 and the other half were on Task 6. As in the stages of subjective holistic scoring and FCE scoring, I typed up the chosen scripts to avoid any effect of the handwriting variable on their assessment, and asked the raters to assess these twenty selected scripts and keep a diary of their rating process.

I wrote a manual on the rating scale to help the raters understand and use it. The manual included an introduction to the scale, a glossary and typical scripts for each band according to the RS1. I also devised a questionnaire (Questionnaire III in Appendix 10) to find out how they felt about the rating scale. This was to obtain feedback on the scale, which would be taken into account when it was revised. I asked them to have a look at the questionnaire before they started the assessment to remind themselves of the aspects that they would need to give feedback on after assessment.

To sum up: the raters were given the rating scale and the manual and asked to assess Group H and make a diary of their rating process. They were also asked to answer the questions in Questionnaire III after using the RS1. I will discuss the analysis results of the diary and the questionnaire in the next section.

### 7.6.3 Feedback on the RS1

#### 7.6.3.1 Feedback from diary

I analysed the obtained diary entries using the same analysis methods as those employed in the previous phase (see section 6.2.3.1 for more details). I summarised the derived findings under the following three headings: rating behaviour, response to the rating scale, and points which need to be revised.

---

[65] Since the scripts on Tasks 5 and 6 were not rated by the raters according to subjective holistic scoring method, I rated Group H by myself to select various levels of script before having them assessed by the raters with the RS1.

#### 7.6.3.1.1 Rating behaviour

The first finding was that the raters still had their own subjective criteria, just as with the FCE scoring. They relied on their subjective criteria in various ways. As noted in sections 6.2.3.2 and 6.3.5.3, Rater A focused on quantity or length when doing subjective holistic scoring (and FCE scoring too), and continued relying on quantity, in particular for extremely low level scripts, as shown in the following diary entry:[66]

> ID: 5-A0105A
> Band: Accuracy-Band 1, Fluency-Band 1, Organisation-Band 1
> This script is too short to assess. I assigned it to the lowest bands for every category without even having a closer look at it. However, I am not sure whether this kind of rating pattern is right.

The raters also relied on their subjective criteria in order to compare the ratings from their own subjective criteria with those from the rating scale.

> ID: 6-A0101A
> Band: Accuracy-Band 4, Fluency-Band 5, Organisation-Band 5
> ...(*the first part of this diary is omitted*) therefore I assigned Band 5 to this script for Organisation. But I am not sure whether it deserves Band 5. It looks to me as though it matches Band 4 according to my subjective criteria. Even so, I assigned Band 5 according to the rating scale.

Their subjective criteria sometimes led the raters to understand the RS1 rather differently from the way I had intended it to be used. For example, Rater A took the top level of the scale (Band 6) as the perfect level of English command, as he did in subjective holistic scoring (see one of his diary entries in section 6.2.3.3, for ID: 3-A0426A). The following demonstrates this point:

> ID: 5-A0108A
> Band: Accuracy-Band 5, Fluency-Band 6, Organisation-Band 6
> It appears that the writer of this script commands fluent expressions and has mastered English grammar. Unfortunately, however, not all the sentences in it are perfect enough for Band 6. These grammatical errors make me hesitant about assigning Band 6. So I assigned Band 5 to this script for Accuracy....(*the rest of this diary entry is omitted*).

---

[66] Notably, however, as revealed in the diaries and verbal protocols in Chapters Seven and Eight, when assessing scripts of other levels using the RS2, Rater A also focused on other features suggested in the scale rather than on length only. It appears that it was generally with extremely low level scripts that he relied exclusively on length for assessing a script.

Second, they still considered the ratings of preceding scripts when assessing a script, as with subjective holistic scoring methods and FCE scoring. This leads me to suppose that this may be common behaviour among raters across different rating methods. This type of behaviour is seen in the following diary entry:

> ID: 6-A0105C
> Band: Fluency-Band 2, Organisation-Band 2
> This script looks extremely poor compared with the previous ones written by other students....(*the rest of this diary entry is omitted*)

#### 7.6.3.1.2 Raters' opinions of the rating scale

The main finding with regard to raters' opinions of the scale was that some of them felt that the differences between the bands were not equal, especially between Bands 2 and 3.[67] This is seen in a diary entry as follows:

> ID: 6-A0103B
> Band: Accuracy-Band 3, Fluency-Band 3, Organisation-Band 6
> It seems to me that the difference between Bands 2 and 3 for Accuracy is bigger than that between other bands...(*the rest of this diary entry is omitted*).

This is not limited to this rating scale and these raters, but can be observed in other rating scales and raters. Raters often see the intervals between the bands of a rating scale, as in (1) below, as unequal, as can be seen in (2) and (3).

(1) 1 2 3 4 5 6
(2) 1 2 3 4 5 6
(3) 1 2 3 4 5 6

(McNamara, 1996: 124)

I did not mean to develop an internal scale such as (3). Even so, as McNamara (1996) points out, the RS1 would have been interpreted differently by the raters.

#### 7.6.3.1.3 Points in the scale which need to be revised

First, the raters reported that the quantifiers and frequency adverbs did not look clear

[67] In fact, the scale was not intended to have interval distances between the bands, but to describe the features of each band on the basis of the scripts.

to them, especially for the differentiation between Bands 2 and 3, and Bands 3 and 4. One rater felt that using the terms for band differentiation was not helpful, as shown in the diary entry below:

> ID: 6-A0103B
> Band: Accuracy-Band 3, Fluency-Band 3, Organisation-Band 3
> ...(*the first part of this diary entry is omitted*) so far using this rating scale I have felt that the differences between Bands 2, 3 and 4 mainly lie in the differences of error frequency, and that interpretation of the frequency adverbs is likely to vary amongst the raters. This is because the concepts of these frequency adverbs and quantifiers are not definite. Since the terms can be interpreted differently by each rater, I am concerned that it could affect the consistency and objectivity of rating.

As a result, whenever they came across these unclear descriptors, they tended to appeal to their subjective judgement to differentiate between bands. In light of this, either the quantifiers and frequency adverbs need to be replaced with clearer terms, or the raters need to be informed of the differences between them through a revised manual for the rating scale.

Second, one of the descriptors needed to be revised as it created a misunderstanding on the part of the raters. This was the descriptor "However, able to write more than required length" in Band 4 of Fluency. It was actually intended to mean that a script was as long as the minimum length, but the raters understood it as "having the potential to write a script of minimum length", as can be seen in the following diary entries:

> ID: 5-A0102C
> Band: Accuracy-Band 4, Fluency-Band 4, Organisation-Band 5
> ...(*the first part of this diary entry is omitted*) Since the writer seemed to have the ability to write more than the minimum length, I assigned Band 4 for Fluency...(*the rest of this diary entry is omitted*).

> ID: 5-A0102A
> Band: Accuracy-Band5, Fluency-Band 3, Organisation-Band 4
> ...(*the first part of this diary entry is omitted*) With regard to Fluency, this script is the minimum length, but I am not sure whether the writer has the ability to write more than the minimum length. So, this aspect does not help me to decide between Bands 3 and 4....(*the rest of this diary entry is omitted*).

This meant that they considered whether the writer appeared to have the

potential to write more if the script was shorter than the minimum length (i.e. less than two hundred words). Therefore, the descriptor needed to be worded differently.

Having considered the findings from the raters' diaries regarding the RS1, I turned to their answers to Questionnaire III, which will be discussed in the next section.

### 7.6.3.2 Feedback from the questionnaire

In addition to the diaries, I obtained feedback from the raters' answers to Questionnaire III. Their first impression of the RS1 was that it was rather complicated. Rater B reported that she spent one hour trying to understand, Rater A took a couple of hours and Rater C needed several hours, as he found it extremely difficult to grasp the rating scale before starting to use it. However, Raters A and B said that this implied that the rating scale was elaborate, and that once they understood the rating scale, it only took them five minutes to rate each script. Rater C required twenty-five to thirty minutes for each script. Therefore, Rater A suggested that rater training would be helpful to get raters accustomed to the scale and help them fully understand it.

There was also some disagreement about certain features in the rating scale. Rater A suggested that the descriptors regarding 'quantity' belong to Organisation rather than Fluency, and that those regarding 'advanced language' and 'advanced grammar' belong to Accuracy rather than Fluency. No doubt these suggestions had some merit, but I explained the rationale for the structure to them in more detail, according to the concepts and definitions of the three main categories and these subfeatures, as outlined in sections 7.3.2.1 to 7.3.2.3 in this chapter.

As I discussed in these sections, 'quantity' can be considered to be one of the subtypes of fluency (Fillmore, 1979), and has been used as a measure of fluency in many previous studies (see section 7.3.2.2).

Additionally, 'advanced language' and 'advanced grammar' were considered to belong to Fluency rather than Accuracy in this study. Since these features deal with grammaticality and language at intra-sentence level, they might appear to be included in Accuracy. However, lack of these features in a sentence or a script would not make it ungrammatical. In other words, a script should not receive a cut in marks

because it does not contain these "advanced" features. Therefore, 'advanced language' and 'advanced grammar' are features whose presence will gain a script "extra" credits that will not be awarded to scripts that do not contain them. Thus, these two features are a matter of language level, of fluent-or-not rather than accurate-or-not, and I still included them in Fluency rather than Accuracy.

With regard to the question of whether anything in the rating scale is hard to understand, Rater A gave a negative answer and Raters B and C an affirmative one. Rater B questioned the method of calculating a final band and Rater C mentioned that he found the differences between bands unclear. Rater B's comment arose because I did not suggest any method of calculating a final /total band after rating for each of the three main categories, and such a method was needed. With regard to Rater C's comment, it appeared that the wording of the descriptors needed to be revised to make the differences between bands clearer, but the content of the descriptors derived from the statistical analysis was not changed.

On the question of whether they found any parts of the scale unnecessary, Raters A and C gave affirmative answers. They felt that the wording in the scale appeared repetitive or contradictory, and needed some revision.

Regarding the question as to which aspect of the rating scale should be maintained, all three raters chose the point that the rating scale had three categories, namely, Accuracy, Fluency and Organisation. They found these categories easy to deal with, with clear, separate indicators of what they were to assess. They mentioned that compared with the FCE rating scale, which included assessment categories that they had not understood, they preferred this aspect of this scale. Rater B added that the concrete indications and explanations under the descriptors were very helpful. She suggested that these explanations be developed into a user's manual which would include concrete and specific examples for each descriptor.

Finally, under 'other comments', Rater C said that it would also be good if the rating scale were developed by an *a priori* method, with the differentiation between bands and assessment categories made on the basis of his experience as a teacher. From his diary entries and answers to the questionnaire, he seemed generally unwilling to accept this type of scale, tending to stick to his own criteria during the assessment. On the other hand, Rater A tried to accommodate the rating scale, and

suggested rater training to help ensure that the rating scale was used as intended.

To sum up, I obtained feedback on the rating scale through the diary entries and questionnaire studies. The rating scale was judged to be good in that it dealt with three clear, independent and simple assessment categories, and tried to give detailed and concrete explanations for descriptors. However, the wording of the scale was thought to be too complicated. The raters felt that it appeared repetitive or even contradictory, and that it led to confusion and hesitancy in their interpretation of the quantifiers and frequency words in the scale, which meant that they had to rely on their subjective judgement. Based on these findings, I tried to revise the scale.

## 7.7 The revision of the RS1

The RS1 was revised on the basis of the feedback from the raters. I tried to improve the wording, suggest how to calculate the total band and revise the user's manual. However, it was not easy to remove the quantifiers from the scale. As the scripts varied in length I could not pick up exact numbers or words for the occurrence of errors to replace the quantifiers and frequency adverbs. Furthermore, using a rating scale does not mean complete elimination of the rater's subjective judgement from the rating process (this is impossible, as discussed in previous literature and revealed in both Chapter Six and the present chapter), and the rating process does not involve mechanically matching the number of errors or occurrence of some features to descriptors in the scale (see section 8.4.8 for further discussion). Therefore, I decided to continue to employ the quantifiers and to explain what they meant in the user's manual more clearly. As a result, the RS1 was revised to the RS2, which is shown in Table 7.10 below.

As can be seen, the RS2 had three findings of information in each cell (i.e., each combination of band and category): an overall idea of the cell; the distinctive features of the neighbouring cell above; and the distinctive features of the neighbouring cell below. The overall idea of each cell was formulated by summarising all the statistical findings of the cell, and was intended to help the raters grasp the general features of the cell. The distinctive features were intended to help the raters decide when they would try to choose which band to allocate the script to, which usually involved

deciding between two candidate bands, as discussed in section 6.2.3.2. I made these general features and the features distinguishing a band from its lower or higher neighbour visually clear, with general features on the top of each cell and distinctive features indicated by arrows, such as '▲' and '▼' before each descriptor.

In addition to the RS2, I revised the user's manual to make it user-friendlier by including detailed explanations and concrete examples for each descriptor, as well as typical scripts for each band, as with the RS1. The manual also gave instructions on how to calculate the total bands from the three scores on Accuracy, Fluency and Organisation.[68]

Table 7.10 The RS2

| | Accuracy | Fluency | Organisation |
|---|---|---|---|
| Band 6 | **Very accurate with few grammatical errors, and better than this**<br>▼few grammatical errors and just a few, if any, local grammatical errors<br>▼No phrase or clause which is unintelligible<br>▼No phrase or clause which has problems in construction<br>▼Or more accurate than these | **Rich and English-like expressions, smooth flow between sentences**<br>▼No case which is incoherent due to irrelevant content<br>▼Smooth flow between sentences due to good command of appropriate clause construction for an intended meaning<br>▼Good clause construction and expansion for rich expression<br>▼Advanced level of language using advanced grammatical features (e.g., aspect, relative adverbs) and advanced clause construction<br>▼Or more fluent than these | **Very well organised in terms of content, structure and paragraphing, good topic address and good development of all the points in the script**<br>▼Fairly good development of all the points in the opening, body and closing<br>▼Or better organised than these |
| Band 5 | **Fairly good and accurate, but with some grammatical errors**<br>▲Some errors across various kinds of grammatical features and a few phrases and clauses which have problems in construction<br>▲Few clauses which are unintelligible due to wrong choice of vocabulary<br>▼No errors with verbs | **Fairly good, but does not command English-like expressions**<br>▲A few local lack of coherence<br>▲No smooth flow between sentences due to failure to use appropriate clause construction for an intended meaning<br>▲No good clause expansion or good expressions<br>▲Little advanced language using aspect and relative adverbs<br>▲Few English-like vocabulary items or phrases<br>▼Good cohesion through little repetitions, use of correct pronouns and connectors beyond those at | **Few problems with content, structure and paragraphing overall, but one (or two) point(s) in the script which is insufficiently developed**<br>▲One (or two) point(s) in the body are insufficiently developed<br>▼Good development in opening and closing,<br>▼No problem with overall structure and paragraphing<br>▼Few errors with paragraphing |

[68] This was done by dividing the three scores for the three categories by three. However, if the scores for both Fluency and Organisation were Band 2, the total/final band for the script was determined as Band 2, without considering Accuracy. Having observed the data, the Band 2 scripts are so short and simple that they have little room to make grammatical errors, to the extent that they could be assigned a high band such as Band 5 for Accuracy. Without this stipulation, therefore, this high band for Accuracy could result in a middle band for a total band which is much higher than real proficiency level from teachers' subjective holistic judgement. Given the ratings done with subjective holistic scoring, where raters refused to put such scripts in a high band, this stipulation was necessary.

| | | | |
|---|---|---|---|
| | | middle school level | |
| Band 4 | **Good, but with a number of literal translations from Korean to English words, and grammatical errors across grammatical features**<br>▲Errors with verbs<br>▲▼Some clauses which are unintelligible due to serious errors in clause construction | **Fairly long, but uses plain English and shows local lack of coherence**<br>▲Plain English in vocabulary and clause construction<br>▲Plain connectors<br>▲Repetition of words without using pronouns and substitution, and makes errors when they are used<br>▼More than minimum length | **Errors with overall structure and paragraphing, more than one point is insufficiently developed, but content and topic address are appropriate and sufficient**<br>▲More than one point in body is insufficiently developed<br>▲Errors with paragraphing<br>▲Errors with overall structure due to unclear closing<br>▼Good development with opening<br>▼Good and clear topic address in opening stage<br>▼Some of the points in the body are insufficiently developed |
| Band 3 | **Sometimes difficult to grasp the meaning due to many words being literally translated into English from Korean and grammatical errors**<br>▲Some clauses which are unintelligible due to either wrong choice of vocabulary or serious errors in clause construction<br>▼Only some clauses are unintelligible | **Around minimum length, plain English, despite frequent local lack of coherence, on the whole manages to sustain overall coherence**<br>▲▼Just around minimum length<br>▼Frequently lacks coherence due to either omission of cohesive devices or irrelevance to the previous sentences in terms of content, but on the whole not disconnected | **Covers the required content but overall structure and development are insufficient**<br>▲Unclear opening and closing parts<br>▲Most points in the body are insufficiently developed<br>▲Sometimes omits topic address<br>▼No paragraphing and, if any, errors with paragraphing<br>▼Covers the required content reasonably well |
| Band 2 | **Quite often difficult to grasp the meaning due to many words, phrases and clauses either being literally translated into English from Korean, or having serious grammatical errors**<br>▲A large number of errors across all kinds of grammatical features including verbs with almost all of the clauses<br>▲Many clauses which are unintelligible due to either wrong choice of vocabulary or serious errors in clause construction<br>▼Follows English word order, i.e. S+V+O/C rather than Korean word order, i.e. S+O/C+V, even though it is very poor with a great number of grammatical errors<br>(When both Fluency and Organisation are judged as Band 2, Accuracy of the script does not need to be judged) | **Substantially less than the minimum length, very plain English, disconnected and generally lacking coherence**<br>▲Less than the minimum length (between 33 and 70% of the minimum length)<br>▲Disconnected clauses across more than 50% of the whole script due to either omission of cohesive devices or irrelevance to the previous clauses in terms of content | **Generally extremely insufficient in both formal structure and content**<br>▲Appears very poorly organised due to no paragraphing and very poor structure<br>▲Opening and closing lacking or unclear, and if any, they are insufficiently developed<br>▲Extremely insufficient number of points in the body<br>▲Extremely insufficient in content |
| **Band 1** | **Less accurate than Band 2**<br>(When both Fluency and Organisation of a script are judged as Band 1, Accuracy of the script does not need to be judged) | **Less fluent than Band 2** | **More poorly organised than Band 2** |

NB. In the scale '▲' indicates features which help distinguish this band from the band above, and '▼' indicates features which help distinguish this band from the band below.

## 7.8 Summary

In this chapter I discussed how the RS1 was developed and revised. First of all, I needed to develop a coding scheme that could be used to code the scripts. I established the three main categories, Accuracy, Fluency and Organisation, on the

basis of the course goals and the definition of writing ability implicit in the English Writing course, and then produced the coding categories by repeatedly reading and preliminarily analysing the writing samples and carrying out a literature review.

Using this coding scheme, I coded all three hundred and ninety scripts. The coding data were then statistically analysed before the features discriminating between neighbouring bands were drawn up. On the basis of the derived distinctive features of the bands, I constructed the RS1 and the user's manual.

The next stage was the trial of the RS1 by the three raters, who were asked to keep a diary for their rating process, and to answer Questionnaire III to determine how they had found the RS1.

The feedback on the RS1 was both positive and negative. The main thrust of it was that the raters found the RS1 rather complicated and felt that the wording needed to be clearer. On the other hand, they liked the fact that it had three clear assessment categories and concrete explanations. On the basis of these comments, the RS1 was revised to produce the revised version, the RS2.

In the next chapter I will discuss the process and the results of the investigation into its qualities as an assessment tool, with regard to its practicality, reliability and validity.

# CHAPTER EIGHT. EMPIRICAL VALIDATION OF THE RATING SCALE

## 8.1 Introduction

In Chapter Seven I discussed how I analysed the obtained writing samples, developed the RS1 and revised it to produce the revised version, the RS2. The RS2 will now be evaluated for the qualities which are traditionally considered important in the language testing field: practicality, reliability and validity. It will also be compared with the FCE rating scale to examine how they differ from each other, and thus whether a need for the RS2 has been established.

I will investigate the rating procedure in section 8.2, its practicality in section 8.3, and its reliability and validity in section 8.4. The RS2 will be compared with the FCE scale in section 8.5, and my summary and conclusion will form section 8.6.

## 8.2 Rating procedure

For this stage, the three raters used the RS2 to assess the scripts of Groups A, B and C (for Raters A, B and C respectively). As in the previous phases, they were also asked to keep a diary on their rating process. About two weeks after completing the assessment, all raters were invited to rate Group D for the investigation of inter-rater reliability and validity of the RS2, and Raters A, B and C were respectively asked to rate Groups E, F and G for the investigation of intra-rater reliability, as in Phase One. The raters were also asked to do a think-aloud during the assessment of the specific six scripts in Group D (see Table 5.1). They did the think-aloud alone consulting the instructions provided by the researcher, recording themselves on cassette tapes according to my instructions and the procedure followed in Phase One (see section 5.6.1 for a detailed explanation of this phase). All data from this procedure were used to investigate the qualities of the RS2.

## 8.3 Practicality

### 8.3.1 Procedure

There are few studies on the practicality of either tests or rating scales, which is usually dealt with only briefly in books on testing. For example, Harris (1969) discusses practicality in his book *Testing English as a Second Language* in terms of

three points: economy, ease of administration and scoring, and ease of interpretation. The first point, economy, has to do with the issues of money and time – how much a test costs to buy, and how many personnel and how much time are required to administer and score the test in the case of purchasing a standardised test. The second point, ease of administration and scoring, is concerned with issues such as whether full and clear directions are provided so that test administrators can perform their tasks properly, whether the test requires elaborate mechanical equipment which may not be available, and whether it is scored subjectively or objectively. The third point, ease of interpretation, includes issues such as whether a test manual, in the case of a standardised test, is provided to help interpret the score easily. Discussing economy as one of the characteristics of a good test, Caroll (1980) explains it as "[whether] the tests provide as much information as is required with the minimum expenditure of time, effort and resources" (p. 16).

These arguments about the practicality of a test can also be applied to the practicality of a rating scale. For example, a rating scale should be clear and easy to understand so that it requires as little time as possible to understand and apply. It is desirable that it needs little time, money and effort to develop and is simple to apply.

The practicality of the RS2 was investigated through a questionnaire that I developed (see Questionnaire IV in Appendix 11), which was designed to cover these three main aspects of practicality.[69] The questionnaire included eight questions, with Questions 2 and 3 on economy, Questions 4, 5 and 6 on ease of interpretation and Questions 7 and 8 on ease of administration and scoring. Question 1 asked for general impressions of the rating scale, and Question 9 what raters found it to use. The survey was carried out in August 2005, after the raters had finished rating the scripts with the RS2. The results will be discussed in the next section.

### 8.3.2 The results

The answers to Question 1, which asked for the raters' overall impressions of the rating scale, were mixed. Raters B and C noted that the RS2 looked trimmed down compared with the RS1, but Raters A and C felt that it still did not look simple, and

[69] One of the three aspects, economy might be investigated by calculating the costs of developing, understanding and administering the scale. However, since it was not feasible for this study, I did not do it but examined the time required to understand and administer it.

found some of the differences between bands ambiguous, so they gave negative overall responses.

Second, with regard to economy, I asked how long it took them to understand the RS2 and how long on average it took for them to assess a script using it. Raters A, B and C answered respectively that it took them five hours, thirty minutes and two hours to understand the RS2, and ten to fifteen minutes, five minutes and thirty minutes to assess a script with it. This implies that Rater B had little difficulty in understanding and using it, that Rater A had some difficulty with it and that Rater C found it hardest to use of the three raters. Rater A could assess a script within fifteen minutes, but required as long as five hours to understand it, while Rater C needed half an hour to assess a script, which does not appear to be economical.

Third, with regard to ease of interpretation, Questions 4, 5 and 6 asked whether there were any areas that they had difficulty understanding, what these were and what they thought the reason for the difficulty was. For this aspect, Rater B answered that none of the RS2 was difficult to interpret, while Raters A and C wrote that they had experienced some difficulty. Rater A found it hard to grasp the differences between bands, citing the difference in Accuracy between Bands 4 and 5 and between Bands 3 and 4, and the difference in Fluency between Bands 3 and 4. He reported that because he had difficulty in understanding these differences, he tended to focus on only one or two features discriminating between two bands. On the other hand, Rater C commented that he had difficulty in interpreting the RS2 because the descriptors and characteristics of the rating scale did not coincide with his personal opinion and view on rating scales for writing assessment, as revealed in part of his comment on Question 6:

> .... (*the first part of his answer to this question is omitted*) Band 5 of Fluency reads, "does not command English-like expressions well". However, I think that it is not desirable to allow a script with not very good command of English-like expressions to be assigned Band 5. That kind of descriptor SHOULD be for Band 4 or lower bands, rather than for Band 5··· (*the rest of his answer is omitted*)

Fourth, for ease of administration and scoring, I asked whether there were any aspects which made it difficult to administer scoring using the RS2. Rater B answered that she had not had any problems in this respect. However, this was not

the case with Raters A and C. Rater A wrote that because he had difficulty understanding certain aspects of the differences between bands, it was hard to administer the scoring; while Rater C reported that he found it difficult to remember all the differences between pairs of neighbouring bands, which meant that he had to consult the scale every time he assessed a script, making it hard to use.

Finally, the answers regarding the overall practicality of the RS2 were mixed. Rates A and B were generally positive. Rater B wrote that although it was questionable whether the RS2 would be appropriate for students at general academic high schools, whose level of English writing is generally lower than the level that the RS2 appeared to deal with, on the whole she found it practical for students at FLHSK. Rater A thought that raters would need to be trained before it could be put to practical use, and addressed its practicality from a rather different viewpoint, saying that it appeared to be practical anyway since there was no other rating scale for students. Rater C questioned its practicality, on the grounds that it took more time to assess using the scale than using subjective holistic scoring method.

The responses to the questionnaire can be summarised as follows: although Rater A needed more time and effort to understand it than Rater B, and had some difficulty administering the scale because he found the differences between bands unclear, Raters A and B generally considered the scale practical to use, whilst Rater C tended to question it. Looking at their views on the practicality of the scale in more detail, it can be seen that Rater B took very little time to understand the rating scale and assess a script using it, and did not have any difficulty in interpreting or using it to administer assessment. However, Rater A needed more time than Rater B to understand and assess a script using it, and Rater C not only needed more time than Rater A, but was generally unwilling to accept the scale, claiming that it would be desirable to change it. It appears that these differences between the raters may be related to their personal background, such as the career history as a teacher, given that Rater C who worked as a teacher for more than thirty years seemed to be accustomed to his own criteria and be unwilling to be open to this new scale, whilst Rater B, who had a only several years of career as a teacher, did not. In addition to this, raters' different views may have to do with insufficiency of training to familiarise them with the scale.

Thus, the practicality of the RS2 in terms of these three aspects is negative with some aspects, indicating that it would be desirable to trim down the scale and formulate clearer descriptors, but is positive with regard to many other aspects.

## 8.4 Validity

### 8.4.1 Procedure for the empirical investigation

Bearing in mind the past and present definitions of validity and validation methods discussed in section 3.3.1.2, and the threat to validity, i.e., construct under-representation or construct irrelevant variance (Fulcher, 1999), I determined to investigate the validity of the RS2 in terms of construct validity on the basis of various kinds of evidence suggested in the literature (Bachman, 1990; Chapelle, 1998, 1999; Fulcher, 2003). In view of the practical limitations, the following six methods of validation methods discussed in section 3.3.1.2 were chosen for this study: correlational evidence, a "G-study" (ANOVA), a MTMM study, two questionnaire studies, diary analysis and think-aloud protocol analysis. The first three methods are quantitative, focusing on the ratings, whilst the others are qualitative methods that look at the rating process. I thereby intended to investigate quantitatively and qualitatively the validity of the RS2.

### 8.4.2 Correlational evidence

For this study, correlational evidence can be gathered in terms of both inter-rater reliability and intra-rater reliability. However, since inter-rater reliability is going to be covered in the MTMM approach in section 8.4.3, the correlational evidence will focus on intra-rater reliability only.

For this analysis, the ratings of Groups E, F and G by Raters A, B and C respectively (see Table 5.1 in section 5.6.4 for more details on these groups of scripts) were compared with the ratings of the same scripts as Groups E, F and G within Groups A, B and C. The results are presented in Table 8.1 below. As the tables show, intra-rater reliability was high in each case and significant at the level of .01.

Table 8.1 Intra-rater reliability using the RS2

| | Rater A1 * Rater A2 | Rater B1 * Rater B2 | Rater C1 * Rater C2 |
|---|---|---|---|
| Intra-rater reliability | .736** | .733** | .821** |
| N | 20 | 20 | 20 |

To recap, intra-rater reliability when assessing according to their own subjective criteria was not significant at the levels of .01, .761 and .780 for Raters A, B and C respectively (see section 6.2.2). In the case of the FCE scoring, there was a slight improvement, compared to the case of subjective holistic scoring, up to non-significance at the levels of .01, .873 and .839 respectively (see section 6.3.4). Compared with those coefficients, the intra-rater reliability of Raters B and C when using the RS2 was as high as in the previous cases. Noticeably, for Rater A, reliability increased significantly up to .736 and his inconsistency in both subjective holistic scoring and the FCE scoring was cleared up. It appears that whilst his subjective criterion and use of the FCE scale were not so firm / consistent, resulting in non-significance of intra-rater reliability, which implies that there is a problem with validity in his assessment, it appears that the RS2 might be of great help to him. In conclusion, these results show that all the raters rated fairly consistently when using the RS2, which implies that from the quantitative approach, the RS2 has content-related evidence of validity.

In the next section, its construct validity will be examined through a "G-study" to find out whether the ratings were influenced by factors other than test-takers' ability.

### 8.4.3 "G-study" (ANOVA)

The ratings of Group D, using the RS2 were employed for the "G-study". Half of them were on Task 3 and the other half on Task 4, which were completed by different writers, as discussed in section 5.5.2.

To begin with, I specified the facets which might affect scores, that is, over which the score should be generalisable. These facets were to be independent variables, while the score was to be dependent variables in ANOVA. For the independent variable, the facet of rater was chosen. The variable of task should, ideally, have been a facet of this study, but it was not possible with the data actually

gathered because the writers who performed Task 3 and those who did Task 4 within Group D were not coincident, as discussed in setion 5.5.2. Consequently, the intended G-study was in fact a one-way ANOVA. The results are presented in Table 8.2 below.

Table 8.2 Results of the "G-study" (ANOVA)

| Source | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Rater | 2.633 | 2 | 1.317 | 1.244 | .296 |
| Residual | 60.350 | | | | |

This table shows that the effect of rater on the test score was not significant. I can conclude, therefore, that the effect of rater on ratings was not significant, and that the score can be generalisable over raters when the raters use the RS2. This implies that the facet of rater does not affect the scores from the use of the RS2 and thus its construct validity.

### 8.4.4 MTMM

For this MTMM approach, the ratings from both the FCE scoring for Group D carried out in Phase One and the RS2 scoring for the same samples at the current phase (Phase Four) were employed. As a result, I gathered the ratings on the same construct measured by different methods (i.e., data for convergent validity investigation) and different constructs measured by the same method (i.e., data for divergent validity investigation). In other words, I obtained the ratings when different methods (i.e., different raters) were employed to assess a construct (i.e., either the construct of the FCE rating scale or that of the RS2) and when the same method (i.e., each of Raters A, B and C) was employed to assess different constructs (i.e., construct of the FCE rating scale and that of the RS2). Since scripts on Task 3 and Task 4 in Group D were produced by different students, as discussed in section 5.5.2, the factor of task was not considered in this study. The results are presented below in Table 8.3.

Table 8.3 Results of an MTMM approach

| | RS2RA | RS2RB | RS2RC | FCERA | FCERB | FCERC |
|---|---|---|---|---|---|---|
| RS2RA | 1.000 | .817** | .890** | .662** | .697** | .791** |
| RS2RB | .817** | 1.000 | .781** | .503* | .724** | .715** |
| RS2RC | .890** | .781** | 1.000 | .551* | .702** | .727** |
| FCERA | .662** | .503* | .551* | 1.000 | .729** | .665** |
| FCERB | .697** | .724** | .702** | .729** | 1.000 | .665** |
| FCERC | .791** | .715** | .727** | .665** | .665** | 1.000 |

NB. 'RS2' means 'using the RS2', 'FCE' means 'using the FCE scale', and 'RA', 'RB' and 'RC' mean 'Rater A', 'Rater B' and 'Rater C' respectively.

I will try to make a judgement on the multitrait-multimethod matrix according to four criteria set out by Campbell and Fiske (1959), reviewed in Fulcher (2003). First, the first criterion is that monotrait-heteromethod correlations should themselves be significantly different from zero. That is, when the same construct (i.e., the RS2 or the FCE) is measured by different methods (i.e., different raters), it should not be zero (i.e., no relationship), but significantly different from it. This could be convergent evidence of the validity of each construct. The coefficients within the enclosed triangles of the table are .729, .665 and .665 for the FCE and .817, .890 and .781 for the RS2. Therefore, this shows that each of these two constructs has convergent evidence of validity. However, it should be noted that the correlations for the RS2 are higher than those for the FCE, which means that the RS2 is much better than the FCE in terms of this aspect.

Secondly, the relationship between monotrait-heteromethod and heterotrait-heteromethod needs to be examined. According to Campbell and Fiske (1959), the former should be higher than the latter. This means that when a construct (i.e., the RS2) is measured by different methods (i.e., different raters), the correlations should be higher than when different constructs (i.e., the RS2 and the FCE) are measured by different methods (i.e., different raters). This is one way of finding discriminant evidence of validity. As found in the table above, the correlations of monotrait-heteromethod, i.e., .817, .890 and .781 within the enclosed triangles at top left of the table are obviously higher than those of heterotrait-heteromethod, i.e., .697, .791, .715, .503, .551 and .702 within the non-enclosed triangles in the table. This indicates discriminant evidence of validity for the RS2. However, since these higher correlations of the former than those of the latter may have been

influenced by the variable of different methods, it is not strong evidence for discriminant evidence of validity. Another investigation needs to be done to find stronger evidence to support the argument.

Thirdly, for stronger evidence, the relationship between monotrait-heteromethod and heterotrait-monomethod needs to be investigated, to determine whether the former is higher than the latter. In the table above, the correlations of monotrait-heteromethod within the enclosed triangles at top left of the table (i.e., .817, .890 and .781) are explicitly higher than those of heterotrait-monomethod in the diagonal lines (i.e., .662, .724 and .727), providing strong evidence for discriminant evidence of validity of the RS2.

Finally, the correlations regarding heterotrait (i.e., the RS2 and the FCE) should show the same patterns across monomethod and heteromethod. This is to find out whether the method (i.e., raters) has any effect on ratings. From the table above, heterotrait correlations show a similar pattern across monomethod (i.e., .662, .724 and .727) and heteromethod (i.e., .697, .791 and .715), which means that the method (i.e., raters) had little effect, which is coincident with the results of the "G-study" (ANOVA) in section 8.4.3.

Another point worth mentioning is the relatively high correlations of heterotrait-monomethod in the diagonal lines in the table. The rather high correlations between these different constructs imply that the two constructs are not totally different, but are associated with each other to some extent. Presumably, this may lie in the fact that since both rating scales are assessing the same skill, i.e., writing skill, rather than totally different constructs, such as speaking and reading, they cannot help having common features even though they assess writing ability in terms of different assessment categories (Accuracy, Fluency and Organisation vs. Content, Accuracy, Range, Organisation and Cohesion, Appropriacy of Register and Format and Target Reader). On the other hand, this result could be interpreted in positive, negative and neutral ways. It is positive in that the new scale concurrent criterion-related evidence of construct validity with the FCE rating scale, a recognised scale, and thus in that it is believed to have construct validity. On the other hand, it is negative in that: (1) the new scale is highly correlated with the existing scale, which might mean that there is little point in developing the new scale as the existing scale could be used; and (2)

the construct of the RS2, which was intended to be assessed through the RS2, was not described in the scale as sufficiently distinctive from that of the FCE scale. On the first of these two negative interpretations, I would say that although the RS2 being found to be fairly correlated with the FCE scale, the rationale for the need to develop the RS2 could be still found in other aspects, such as the need for a scale incorporating the characteristics of Korean students and teachers (see sections 1.6.4.3.2 and Chapter Six) in behalf of raters as well as students, the problems of published rating scales (see sections 1.6.4.3.3, 4.5.2 and 6.3.5.4) and the raters' positive opinion of the RS2 over the FCE scale (see section 8.4.5). The second point might indicate an area in which a further study is needed. In addition, neutrally stated, since this high correlation could result from a hidden variable, this does not mean that the two scales measure the same things, as Fulcher (2003) points out about his own similar findings.

In conclusion, the RS2 has both strong convergent and discriminant evidence of validity, compared to the FCE, which is the evidence why the RS2 could be used preferably to the FCE.

### 8.4.5 Questionnaire V

The validity of the RS2 was also investigated through a questionnaire study filled in by the three raters. This was intended to find out whether the ratings using the RS2 could be interpreted as valid, by using only open-ended questions in order to obtain the raters' detailed opinions. It was sent by email at the end of December 2005, and responses were collected by 10th January 2006. The questionnaire consisted of nine questions regarding the validity of the rating scale (see Appendix 12 for the questionnaire). I will discuss these questions and their responses.

First, Questions 1 and 2 were to find out whether the raters understood the RS2 as intended, which would allow me to conclude that their ratings using the RS2 are valid. All the three raters responded that they found the manual helpful in understanding the RS2, and that on the whole they thought they understood it as intended, although they were unsure about aspects requiring subjective judgement, such as English-like expressions. It is, however, noteworthy that, given Cohen (1994) discussed in section 3.3.2 and Hill and Storch (1994) discussed in section 4.4,

assessment of aspects requiring such subjective judgement seem to be hard to other raters, too.

Second, Questions 3 and 8 dealt with the issue of whether the raters' subjective holistic judgement influenced their assessment when they used the RS2, which might negatively affect the validity of the ratings given with this scale. For Question 3, all the raters wrote that they consulted the scale every time they assessed a script, which indicates that they tried to stick to the RS2 and avoid the influence of their own subjective criteria. In addition, for Question 8, Raters B and C replied that they did not consider aspects other than the three main assessment categories of Accuracy, Fluency and Organisation. However, Rater A appears to have added to one category, as revealed in his response below:

> Since it was challenging for me to assess in terms of these three aspects, I did not dare to consider other assessment categories, but I might have considered it if the scripts had been hand written. I would have taken into account whether a script is written neatly or not. However, since all the scripts were typed, it was not the case. The most influential aspect of the rating scale was the length or quantity of a script. It was included as one of the features in Fluency. However, it was more than just a feature to me. I have to admit that it was the most powerful aspect for me. If a script was very long, it looked as if it deserved extra credit, but if it was too short, it looked superficially as if it was not well done. So it seems that I took more account of length than I should have done.

Given that the influence of subjective criteria on the use of rating scales was also observed in the FCE scoring and the RS1 scoring, it appears that this is likely to happen. Another point to mention with regard to his response above is that it appears to be due to the influence of his subjective criteria which was found to put weight on length that he thought he took more account into length than he should have done. It does not appear to be because the feature of length was devised, unlike my intention, to appear more salient than the other features with all bands in the scale rather than with some bands where length was suggested as one of distinctive features for Fluency. This assumption is supported by the fact that this phenomenon with Rater A was not the case with the other two raters who did not have weighted length in their own subjective criteria. Meantime, given that even Rater A achieved high intra- and inter-rater reliability in the RS2 scoring unlike in both subjective holistic scoring and the FCE scoring, his attention on length may not have been so excessive as he feared.

This will be investigated further in section 8.4.7.

In conclusion, apart from the fact that Rater A "thought" he considered 'length' more than suggested, the raters generally tried to assess only in terms of the assessment aspects in the RS2, and consulted the scale for the assessment of every script.

Third, Question 4 was meant to investigate whether the validity of their ratings varied depending on the genre of scripts. When I developed the RS2, I did not intend the rating process and application of the scale to vary according to the genre of a script. If this was the case, it would have an unintended effect on the validity of the ratings using the RS2, and it would be said that there is a difference in validity of the rating scale depending on the genre of the script being assessed. Rater A responded to Question 4 saying that he applied the scale equally regardless of the genre of a script, while Raters B and C reflected that they tended to apply the scale differently when assessing formal essays and private letters. Both responded that it was easier to use the scale for formal essays than for private letters. Rater C did not give any detailed reason for this, but Rater B wrote:

> I found the scale more applicable to the assessment of formal essays than private letters. Since the structure and issue of topic address required in formal essays are clear, and the English should be formal, it is easier to establish the right model for these aspects in formal essays. That is, it is obvious what a formal essay should look like in terms of Accuracy, Fluency and Organisation. On the other hand, however, in the case of private letters, other than the requirement that they need lexical expressions such as "*Dear ~*", "*Hello*", "*Love*" or "*Best regards*" and a signature, I could not find any concrete criteria for their organisation or structure, so I did not find the scale very applicable to private letters. Furthermore, since I think colloquial expressions should be allowed to some extent in private letters, it seemed to me that to strictly apply the rule of grammar for the assessment in Accuracy might not coincide with the situation in the real world. However, given that the scripts were written for assessment, I thought that they should be formal. But since I could not decide how strict I should be for the grammar of the written language, I didn't find it easy to apply the scale to private letters.

In summary, according to Rater B, since the structure and English for formal essays should be evidently formal, it was easy to establish the ideal model of formal essays in one's mind, and consequently it was also clear how to judge whether a script matched the standard or not. On the other hand, since she was not given any concrete guidelines determining how colloquial the English in private letters could be

in a test context, she was hesitant about applying the RS2. In light of this, I can conclude that raters need to be provided with relevant guidelines to enable them to assess such scripts using the RS2 as intended.

Fourth, Question 5 dealt with the advantages and disadvantages of using the RS2 rather than subjective holistic assessment. I devised this question in case their answers might be concerned with validity of the RS2. For this question, Rater A mentioned that the advantage of using the RS2 was that it helped increase the reliability of rating. Rater C answered that he mainly considered grammatical accuracy in his subjective holistic assessment, and that it was good to take account of the different aspects covered by the RS2. Rater B's response was more evidently concerned with validity, as follows:

> It seems to me that it took less time to assess using the RS2 than according to my subjective criteria. At first the latter might appear to proceed more quickly than the former, but as I went on assessing according to my subjective criteria I had to go back and check the ratings I'd given to previous scripts because my criteria became vague and confusing in terms of assessment categories as well as band differentiation, and I became unsure of them. Although it took time to understand the RS2 at first, it was easier once I'd grasped it because the assessment categories and band differentiation are clearly established to help keep me on the right track, and this was a definite advantage of using a scale…(*the rest of her response to this question is omitted*)

In brief, the raters responded that the scale helped them with regard to which features they should assess, as well as making the rating more consistent.

Fifth, in Question 6, I asked whether they thought the RS2 could provide a valid picture of Korean students' writing ability. All of the raters answered affirmatively, although they had different reasons for their answers. Rater A said that it was because the scale was developed on the basis of writing samples from Korean students, and because it looked good in terms of its content as well as form. However, he added that the effectiveness of the RS2 would depend on whether teachers are well acquainted with it and whether they could apply it as well as they understood it. Rater B revealed that she trusted in rating scales, saying that compared with subjective holistic assessment, using the rating scale in its own right could provide a valid picture of learners' writing ability. Rater C commented on this point that it was because assessment was conducted in terms of various aspects.

Sixth, Questions 7 and 9 asked whether the raters thought that the RS2 assessed the three aspects of Accuracy, Fluency and Organisation properly, what they thought the backwash effect of the RS2 would be, and what the differences between the RS2 and the FCE rating scale were. Raters B and C responded that the RS2 appeared to assess the three assessment categories properly. Rater A, however, differed slightly from them, pointing out that fluency was not as well assessed in Fluency as accuracy and organisation were in Accuracy and Organisation respectively. His remark about this point is as follows:

> ...Since Accuracy deals with grammatical correctness on the whole and Organisation does structure and paragraphing, accuracy and organisation of a script appear to be assessed in these two categories. The features in Fluency are theoretically appropriate for this category. However, the problem is that the features in Fluency such as 'coherence', 'smooth flow between sentences due to good command of appropriate clause construction for an intended meaning' and 'good clause expansion and rich expression' were so ambiguous that I found it hard to assess any of these features well except for 'length' in Fluency. I mean, whilst Fluency theoretically assesses fluency properly in that it includes proper features in the assessment, it does not do its job properly in practical terms because these features are not clear to assess.

It is supposed that since some features in Fluency required subjective judgement, this made him find the assessment category hard and unclear to assess, as discussed above. However, all three raters agreed that using the RS2 would have a positive backwash effect of leading teachers and students to pay attention to the features in the scale, which should be considered in the teaching and assessing of writing in their writing course, and that the RS2 was found to be different from the FCE scale in that the former had more concrete descriptors and assessment categories and fewer assessment categories than the latter. This may have helped the raters understand the descriptors and assessment categories, and as a result improved the validity of the scale.

I have reviewed the raters' responses to Questionnaire V, which investigated the validity of the RS2. Whilst there were some negative responses, there were many positive ones as well: the raters thought that on the whole they understood it properly as intended; they tried to stick to it during the assessment; they thought that it helped them stay on the right track and be more consistent than with subjective holistic scoring; they thought that it would provide a valid picture of Korean students'

writing ability; and they agreed that it would have a positive backwash effect on the teaching and learning of English writing. Therefore, it can be concluded that the raters found it valid on the whole, and that their ratings using it were valid. However, the validity of this RS2 could be further improved by revising the features in Fluency to make them clearer, and by providing the raters with more concrete information on the assessment of private letters in a test context.

### 8.4.6 Diary analysis

The diaries kept by the raters when they were using the RS2 to assess Groups A, B and C were collected for analysis for the investigation into the validity of the RS2. As with the previous diary studies, following the same analysis procedure (see section 6.2.3.1) this entailed investigating both the general patterns and tendencies in their use of the RS2, and the specific features of the scale which the raters focused on. I categorised the points revealed under the following five headings regarding validity: the raters' mode of application of the RS2 which may have affected validity of the RS2; the features in the scale mentioned by the raters; the aspects which appeared to be either tricky/ambiguous or particularly easy to deal with in the RS2; and the aspects which raters perceived differently from the way suggested. I will discuss each point in detail.

First, it seemed helpful to investigate how the raters used the RS2 in order to examine its validity. This point is divided into two sub-points: the consideration of subjective holistic criteria, and non-adherence to category divisions. Regarding the first point, the three raters considered their subjective criteria to some extent while using the RS2, as in previous scorings. This was found in various aspects. It was observed with regard to the category for which they assessed the scripts, as they not only assessed them for categories which were included in the RS2, but also sometimes considered other assessment categories that were not included in it, such as the issue of symmetry between paragraphs in the degree of development (for Raters A and B) and the issue of whether the writer had tried as hard as they could (for Rater C). As in his subjective holistic scoring, Rater C also wrote that more weight should be given to Accuracy rather than Fluency and Organisation, as shown in the extract from his diaries below.

ID: 3-B1101C
Band: Fluency-Band 2, Organisation-Band 2, Final-Band 2
...(*the first part of this diary is omitted*) For Fluency, given that this script is extremely short and written in very plain English, that coherence is poor on the whole, and **since it seems that the writer has not done his/her best**, I gave Band 2 to it...(*the rest of this diary is omitted*)

ID: 4-B1111C
Band: Fluency-Band 2, Organisation-Band 2, Final-Band 2
This script has many errors with grammatical categories which hinder intelligibility, and it even has many errors in sentence construction. For example, this writer writes, *can't help but asking* rather than either *cannot help asking* or *cannot but ask*. In addition, a compulsory complement, i.e. an object after *only one click* is missing, which makes the sentence unintelligible. Since there are many awkward expressions in it, I'll give it Band 3 for Accuracy. For Fluency, it's long enough, but the overall fluency is not good. For example, given that the writer omitted *I am* in a sentence *What a pity person I am* and *we are* in a sentence *What pity persons we are*, I'd say the fluency of this script is extremely poor. And there are clauses which omitted objects or conjunctions. So, Band 2 seems suitable for Fluency. For Organisation, since Closing is insufficient and the disadvantage of the use of the Internet (one of the major topics) is omitted, I gave it Band 2 for Organisation. According to the rating scale, in cases where a script is given Band 2 for both Fluency and Organisation, it does not need to be assessed for Accuracy and its final band is Band 2. However, it's a shame to give Band 2 for its final band since it could well have Band 3 for Accuracy. **So I disagree with prioritising Fluency and Organisation or giving them more weight than Accuracy.[70] I think Accuracy should have more weight and priority in assessment than the other two categories.**

The consideration of their subjective criteria was also manifest in the manner of comparing the ratings from the RS2 with those from their subjective holistic assessment, in Rater C's diary:

ID: 4-B1131C
Band: Fluency-Band 2, Organisation-Band 2, Final-Band 2
This script is too short, its opening is extremely insufficient, not well organised and there is no closing part in it. Therefore, it cannot deserve a good band. For both Fluency and Organisation, I gave Band 2. According to the rating scale, Accuracy for this kind of script does not need to be assessed. Therefore, the final band came to be Band 2. **On the other hand, if I just consider English at the sentence level as I do in my subjective holistic assessment, this script would deserve more than Band 2 for its final band. If it is assessed only for Accuracy, it should get Band 4 or so.**

[70] In fact, Fluency and Organisation were not prioritised or given more weight than Accuracy in the scale. The reason why I decided that Accuracy does not need to be assessed when both Fluency and Organisation are put in Band 2 is discussed in the footnote in section 7.7.

To summarise, the raters considered their own subjective criteria when using the RS2, just as they did when they did using the FCE scale and the RS1. They sometimes considered assessment categories which are not included in it, and compared the ratings from subjective criteria with those from the RS2, as they did for FCE scoring (see section 6.3.5.3 for more details). In light of this, I can safely conclude that raters are likely to consider their own subjective criteria when using any kind of rating scale.

For the second subpoint, analysis of the diaries revealed that Rater C did not always adhere to the category divisions. His diary shows that it did sometimes occur. When he was assessing Fluency, he also assessed 'content', 'format', 'structure', 'spelling', 'grammatical errors' or 'intelligibility of sentences' that were to be dealt with in either Accuracy or Organisation. It can be seen from this that he did not always adhere to the category division suggested in the scale. The following diary extracts are examples of this.

> ID: 3-B1102C
> Band: Accuracy-Band 4, Fluency-Band 4, Organisation-Band 4, Final-Band 4
> …(*the first part of this diary is omitted*) While the writer was trying to explain Kyungbok Palace, he/she used *temple* rather than *palace* and *In the stadium watch the match is a good way and good experience* rather than *watching the match in the stadium is a good way to see it and a good experience*. In addition, **since there are some unintelligible clauses in it** and since its length, language level and grammatical level are plain, **I should give Band 4 for Fluency**…(*the rest of this diary entry is omitted*)

> ID: 3-B1110C
> Band: Accuracy-Band 5, Fluency-Band 3, Organisation-Band 3, Final-Band 4
> …(*the first part of this diary is omitted*) For Fluency – since it's just about the minimum length, its fluency is not good. Coherence is good on the whole, **but as its content is only about Jeju-island, content is insufficient. Therefore, it should be Band 3 for Fluency** …. (*the rest of this diary is omitted*).

This non-adherence to category divisions seems to have been caused by consideration of his own subject criteria. I guess that since Rater C had his own solid criteria developed during a long career of more than thirty years in teaching, he sometimes tended to stick to his own subjective criteria, whilst Raters A and B rarely did.

Second, it seemed helpful to examine which aspects of the features in the RS2

the raters mentioned when investigating its validity. I presumed that the aspects which the raters mentioned would help reveal what they thought the RS2 assessed. Analysis of their diary entries shows the features in the RS2 that the raters mentioned as follows:

Table 8.4 Features mentioned by raters in their diaries while using the RS2

| | Accuracy | Fluency | Organisation |
|---|---|---|---|
| Rater A | 'errors pertaining to verbs'<br>'errors across various grammatical categories'<br>'literally translated words/phrases' | 'length'<br>'the use of pronoun'<br>'the use of connectors'<br>'coherence'<br>'advanced language' (language level) | 'structure of opening and closing'<br>'paragraphing'<br>'development'<br>'content' |
| Rater B | 'errors with regard to either verbs or sentence construction'<br>'intelligibility'<br>'grammatical errors across the grammatical categories' | 'length'<br>'advanced language' (language level)<br>'coherence'<br>'the use of connectors' | 'structure' (opening-body-closing)<br>'paragraphing'<br>'development'<br>'content'<br>'genre format'<br>'topic address' |
| Rater C | 'grammatical errors'<br>'errors with mechanics' | 'length'<br>'advanced language' (language level) | 'content'<br>'structure'<br>'development' |

When the features which they mentioned for each assessment category shown in the table above are compared with the descriptors of the RS2, they appear to understand the essentials of the RS2. This was particularly the case with Accuracy, for all three raters. I suppose that it may have been evident to them because Accuracy has to do with grammatical correctness at sentence level. For Fluency, all three raters focused on 'length' and 'advanced language' (language level), which may be because these two concepts are either clear or common for fluency-related features. Raters A and B mentioned other features as well, whilst Rater C just focused on these two features. For Organisation, Raters A and B focused on 'structure', 'paragraphing', 'development' and 'content', whilst Rater C emphasised 'structure', 'development' and 'content', but not 'paragraphing'.

Although the raters focused on many main features of each category in the RS2, they did not pay attention to every feature in each category. Some features, such as 'topic address' (for Raters A and C) and 'paragraphing' (for Rater C) were not mentioned, even though they were supposed to be dealt with within the general characteristics in each cell of the RS2. It is possible that the raters were unclear about the other features which they tended to leave out, found them unimportant or

disagreed with them.

The raters focused on some features of the RS2 rather than all of it, but those that they did pay attention to generally included its main features. I can conclude, therefore, that although they did not consider as many features for Fluency as they did for Accuracy and Organisation, they took many of the main points of Accuracy and Organisation in the RS2 into consideration and that their ratings using the RS2 are fairly valid. In that these untrained raters interpreted the scale silimarly to each other, paying attention to its essentials and their ratings showed high inter-rater reliability, as shown in section 8.4.4, this is evidence for convergent evidence of validity of this scale.

Third, it seemed helpful to investigate the aspects which raters found tricky/ambiguous or particularly easy to deal with in the RS2 in order to ascertain its validity. They seemed to have most trouble with the quantifiers, frequency adverbs, Fluency and 'coherence' in Fluency. The raters reported in their diaries that since they were not entirely clear about the meaning of the quantifiers and frequency adverbs, they could not make a firm differentiation between bands on the basis of them, as they had pointed out when using the RS1. On the whole they did not find it difficult to pick up errors and insufficiency in Accuracy and Organisation, but tended to have problems assessing the degree of fluency, since the concept of fluency tends to be determined across a whole script and is too abstract to pick out parts which are either fluent or dysfluent. Even Rater B, who tried to cover many features in the RS2, focused on mainly 'length' and 'language level'.

It had been my intention to make 'coherence' an important part of Fluency, but there was little reference to this feature in the diaries, possibly because it was tricky to deal with. I had found this to be so when I tried to establish this feature as one of the coding categories in the coding scheme and to operationalise it for the consistency analysis of this feature. The MSc student in applied linguistics who helped me with quality control in coding data of the coding scheme (see sections 5.6.2 and 7.4 for more details) also commented that the feature was tricky to apply, and I assume that the raters had similar difficulties with the concept, as the diaries below show only a passing mention of it.

ID: 3-A0413A
Band: Accuracy-Band 4, Fluency-Band 5, Organisation-Band 4, Final-Band 4
Paragraphing in this script is not well done. **Its coherence is poor**.
Whilst its opening is good, closing is not so good.
Although its length is absolutely sufficient, its overall quality (level) does not look great. There are even errors in the verbs in it.
Therefore, I assigned Band 4 for Accuracy, Band 5 for Fluency and Band 4 for Organisation.

ID: 3-B0118B
Band: Accuracy-Band 4, Fluency-Band 5, Organisation-Band 6, Final-Band 5
...(*the first part of this diary is omitted*) For Fluency, this writer does not have a very good command of English. **Still, there are not many cases of incoherence in this script**, and it's long enough, so I gave it Band 5.

On the other hand, there were some features that none of the raters mentioned as problematic, such as Accuracy and Organisation. While they may have found it hard to pinpoint errors in Fluency, Accuracy was an entirely different matter. They had no trouble making judgement on errors in grammatical features, and Organisation does not seem to have been so tricky for them, either. It is supposed that this is because Accuracy and Organisation are more concrete than Fluency. The ease with which they dealt with these two categories should be conducive for the validity of the ratings of the RS2.

Finally, it emerged that a feature of the descriptors was not perceived as intended by Rater C. It was 'topic address', which was originally supposed to deal with whether a statement is made in the opening part of the script, informing the reader what it will be about and what its purpose will be. However, from Rater C's diary, it is evident that his perception was quite different. It seems that he understood 'topic address' as about whether the theme of the script matched the given topic, as uncovered in the diary below.

ID: 3-B1102C
Band: Accuracy-Band 4, Fluency-Band 4, Organisation-Band 4, Final-Band 4
Since this script has a fair number of awkward expressions and grammatical errors, Band 4 should be appropriate for it in terms of Accuracy. While the writer was trying to explain Kyungbok Palace, he/she used *temple* rather than *palace* and *In the stadium watch the match is a good way and good experience* rather than *watching the match in the stadium is a good way to see it and a good experience*..
In addition, since there are some unintelligible clauses in it and since its length, language level, and grammatical level are plain, I should give Band 4 for Fluency.
**In addition, although development is not so good, since for topic address,**

**content matches the given topic, I gave Band 4 for Organisation.**

I have discussed five points regarding the inquiry into the validity of the RS2 from the diary analysis. On the whole the raters perceived and used the RS2 as intended. However, partly because their application of the RS2 was still influenced by their subjective criteria, and partly because some features in the RS2 such as 'coherence', Fluency and quantifiers were not clear enough to them, they sometimes used and perceived the RS2 in ways that were not intended. Nonetheless, given that the features that they mentioned during the assessment were coincident with the essentials of the RS2, and the raters were not trained enough, so as to reveal that they could independently interpret and use the scale as intended on the whole, these findings are considered conducive to its overall validity.

### 8.4.7 Think-aloud protocol analysis

As I mentioned in section 8.2, the raters were asked to do a think-aloud in Korean while they were assessing six specific scripts from Group D. This think-aloud study would help determine the validity of the RS2 by revealing: (1) their rating process and (2) which aspects of each assessment category they paid special attention to, in a similar way to the diary study in section 8.4.6. The first point was supposed to reveal how they used the RS2 during the assessment, and the second what they thought Accuracy, Fluency and Organisation would assess. In other words, this would show whether each category had been understood as intended, and thus whether their ratings could be said to be valid.

The first step of this analysis was to transcribe the verbal protocols recorded on the cassette tapes. It was analysed according to the same procedure as in section 6.2.3.1. The findings from the analysis can be discussed under the following two headings: rating behaviour; and interpretation of the three main assessment categories. First, I will discuss the raters' rating behaviour. It is worth noting that unlike Raters B and C, Rater A often reacted to his first impression of the script under assessment, which mainly focused on the amount of writing. This shows that quantity was a very important feature for him, as demonstrated in the excerpt below.

ID: A-RS2-TA-01[71]
Well.....It looks very short.... Very short...
It is less than half of usual length of other scripts...
(*he starts reading the scripts*)
Hi, ____. I heard you'll visit Korea next week.
.....(*the rest of this protocol is omitted*)

As this excerpt shows, Rater A noted the quantity of the script before he had even started reading it. This appears to coincide with his response to Question 8 in Questionnaire V (see section 8.4.5). As he answered the question, he gave more weight to quantity than suggested in the instructions.

Of the three raters it is noticeable that Rater B referred to the descriptors in the RS2 most often during assessment. This implies that she tried to follow and adhere to the rating scale, as she said in her answer to Question 3 in Questionnaire V (see section 8.4.5). The following excerpt demonstrates this:

ID: B-RS2-TA-02
...(*the first part of this protocol is omitted*)
According to the descriptors of Band 4 in Organisation..... it says that there are errors with overall structure due to unclear closing (or opening) and that either opening or closing is well developed, so these descriptors appear to match this script..
So, Band 4 for Organisation.....
... (*the rest of this protocol is omitted*)

Rater B also differed from Raters A and C in terms of assessing order. Whilst Raters A and C assessed in the order of Accuracy, Fluency and Organisation, Rater B started with Organisation, then assessed Accuracy and finally Fluency. This seems to imply that Rater B, unlike Raters A and C, considered overall structure and organisation relatively important.

Rater C did not always agree with the RS2. He often hesitated over giving a mark based on the RS2, and tended to rely on his subjective criteria. He also compared marks given on the basis of the RS2 with those based on his subjective criteria, which the other two raters did not do. His diary and answers to Questionnaire V suggest that he was more accustomed to using his own subjective criteria than the other two raters, and that he tended to consider his own criteria while

[71] In the excerpt, "RS2" in the Id number means that the rating was done using the RS2.

using the RS2 more than the other raters.[72] An example of his protocol is presented below.

ID: C-RS2-TA-01
... (*the former part of this protocol omitted*)
Looking at Fluency, this script is shorter than the minimum length..
According to the rating scale, I should give it Band 2 for Fluency.
If so, the total band would be Band 2. But since I think this script seems to deserve a higher band than Band 2, I hesitate to give it Band 2 for Fluency....
... (*the rest of this protocol is omitted*)

Second, pertaining to the interpretation of the three main assessment categories, I summarised the categories that each rater mentioned during the assessment in Table 8.5 below, which represents how they interpreted each of the three assessment categories.

Table 8.5 Features mentioned by raters in their verbal protocol while using the RS2

| | Accuracy | Fluency | Organisation |
|---|---|---|---|
| Rater A | 'general linguistic accuracy'<br>'grammatical category' | 'length'<br>'pronoun/repetition'<br>'language level' | 'structure'<br>'paragraphing'<br>'development' |
| Rater B | 'general linguistic accuracy'<br>'grammatical category'<br>'awkwardness' | 'length'<br>'language level' | 'structure'<br>'development' |
| Rater C | 'general linguistic accuracy'<br>'vocabulary: accuracy and usage'<br>'mechanics'<br>'grammatical category'<br>'sentence construction'<br>'awkwardness'<br>'clarity/communication (clause extensions)' | 'vocabulary, phrase: sophistication and appropriateness'<br>'cohesion' | 'overall organisation'<br>'structure'<br>'development'<br>'content' |

These features in the think-aloud protocol are fairly similar to the features mentioned in the diaries, although there are some small differences between them. As can be seen in the table above, Raters A and B interpreted the scale, focusing on a few features within each of the three categories. They focused on reasonably representative features for the category of Accuracy; however, with Organisation they paid attention to 'structure' and 'development' but not to 'content', although they did attend to 'content' in their diaries. It is a similar story with Fluency, where

[72] As discussed in section 8.4.6, the other raters also sometimes considered their own subjective criteria.

they mainly mentioned 'length' and 'language level' and did not remark on 'coherence', although this was mentioned in their diaries. As Weigle (2002) explains (see section 3.3.2), it may be that they narrowed down their range of focus and 'interpreted' the scale in their own way, according to their particular background, as they went through the assessments (diary-keeping was carried our prior to think-aloud).

Despite certain similarities with the other two raters, Rater C was also slightly different from them. He considered 'content' as well as 'structure' and 'development' for Organisation. In addition, he also considered other features such as 'vocabulary and phrase: sophistication and appropriateness' and 'cohesion using connectors' for Fluency, unlike the other two raters, who were more concerned with 'length' and 'language level'. Furthermore, he gave more weight to Accuracy than the other two raters, which was probably a reflection of his subjective criteria, which were mainly concerned with grammatical accuracy at sentence level (see section 6.2.3 for more details). As can be seen in his protocol below, he commented on all grammatical errors spotted in the script and corrected them.

ID: C-RS2-TA-02
… (*the first part of this protocol is omitted*)
For this reason, I've got little frightened, but still, I'm OK.
It seems to me that this sentence is intended to mean that 'I was frightened'. If so, it should have been "I was a little frightened", inserting an indefinite article "A" …
I heard from Teresa that you'll come to Korea in this winter.
This sentence is wrong. It should have been written as "I heard from Teresa that you would come to Korea this winter". I am afraid there are two errors in just one sentence.
I'm so glad to hear that. Finally, we're gonna meet and see our faces, aren't you?
In this sentence since the subject of this sentence is "we", tag question should have been "aren't you" or "aren't we".... Wrong tag question...
Expecting the day, I'd like to recommend to you some place to visit.
In this sentence, "some place to visit" should be corrected to "some places to visit" because a plural noun, "some" does not agree with a singular noun, "place"..
… (*the rest of this protocol is omitted*)

Another point worth noting with regard to the interpretation and application of the scale is the issue of the dominance of 'length' in the raters' assessment. As discussed in the part on Length in section 7.3.2.2, the inclusion of the feature 'length'

could have made it a surrogate for all other features. However, as the protocols above (and the diaries in section 8.4.6) reveal, the raters only mentioned length when necessary, that is, when it was suggested as a distinctive feature for some bands (i.e., Bands 2 vs. 3 and Bands 3 vs. 4), and it did not dominate the other features (except the case of Rater A who tended to rely exclusively on length when assessing extremely low level of scripts using the RS1, as discussed in section 7.6.3.1.1).

To summarise, I examined the raters' rating behaviour and their interpretation of the main assessment categories. The results of the analysis are similar to those of the diary analysis, showing that they sometimes perceived certain categories differently or more simply than intended and they did not consider some features in Fluency and Organisation. Although there were some disparity between their application and what was suggested in the scale and the manual, the raters generally seem to have understood the RS2 properly and tried to adhere to the scale, given that the features which they focused on were usually included in the main features for each category in the scale. I believe that their use and interpretation of this scale like this are fairly positive to the validity of the scale, given that the raters were untrained sufficiently other than getting themselves accustomed to the scale through its manual.

### 8.4.8 Questionnaire VI

Following Fulcher (1996c), the validity of the RS2 was also examined through the questionnaire study for students to whom the RS2 could be applied. The aim of this questionnaire was to determine whether the students would find the RS2 valid.

I constructed Questionnaire VI in the form of a semi-structured questionnaire, consulting the questionnaire in Fulcher (1996c). As can be seen in Appendix 13, Questionnaire VI included four main questions in the form of closed questions: whether the RS2 looked appropriate as a rating scale for assessing Korean students' writing; whether they thought that the RS2 assessed appropriate features for writing assessment; whether they thought the demarcation between bands was done properly; and whether they thought the RS2 would provide a valid picture of their writing ability. For further information on each question, I also created one or two open-ended questions asking them to give the reasons for their answer.

The respondents to the questionnaire were eighty first-year Korean students

from two classes at Kwacheon foreign language high school, thirty-six of whom were male and forty-three female. None of them had been asked to do any writing tasks for this study in 2003 (see section 5.3). Since this questionnaire study was carried out in 2005, and since classes were organised differently every year in Korea, the subject group that had completed the written tasks was no longer available. Instead, the study was conducted on new students in the charge of Rater A in 2005. Since the respondent group did not have to be coincident to the subject group for obtaining feedback on the RS2, this did not matter at all.

The questionnaire was distributed to eighty students, but as one of them did not answer at all, it was left out of the study. Therefore, seventy-nine questionnaires were analysed for this study. The respondents were between fifteen and sixteen years old, and studied English Grammar under Rater A in 2005. The gender and English score of the respondents are shown in Table 8.6 below. As can be seen in the table, more than half of the respondents belonged to the first and second highest-scoring English groups (over a grade of eighty-one).

Table 8.6 Gender and English score of the respondents to Questionnaire VI

| | | English score | | | | Total |
|---|---|---|---|---|---|---|
| | | 51-70 | 71-80 | 81-90 | 91-100 | |
| Gender | Male | 3 (3.8%) | 12 (15.2%) | 16 (20.3%) | 5 (6.3%) | 36 (45.6%) |
| | Female | 0 | 4 (5.1%) | 22 (27.8%) | 17 (21.5%) | 43 (54.4%) |
| Total | | 3 (3.8%) | 16 (20.3%) | 38 (48.1%) | 22 (27.8%) | 79 (100.0%) |

The survey was carried out under Rater A's supervision in December 2005, during one of their English Grammar classes. For this survey, he started by giving every student a sheet of the RS2 and then helped them go over it, reading it aloud and pointing out the key points of its characteristics. He then allowed time for the students to look at it by themselves before distributing the questionnaire sheet and asking them to answer to it.

The seventy-nine respondents answered all the closed questions in the questionnaire, but not all of the open-ended questions. Some students answered every open-ended question, some answered some of them and others did not answer any. Some of the answers to the open-ended questions revealed that the point of the question had not been understood as intended. Having reviewed the answers, I

imagine that the lack of response or inappropriate answers may be due to respondents not being sufficiently motivated to think hard about the RS2 and the questionnaire; not having enough time to fully understand either the RS2 or the questionnaire;[73] and not having either the knowledge or technical terms to express their ideas about language testing or applied linguistics. Consequently, less than seventy-nine answers to each open-ended question were analysed.

The answers to the closed questions in the questionnaire were analysed by an SPSS computer package. The answers to the open-ended questions were assembled and grouped according to category, and the frequency of cases for each category was counted. I will now discuss the analysis results.

First of all, with regard to the question as to whether the RS2 looked appropriate as a rating scale for assessing Korean students' writing, Table 8.7 below shows that 91.1% of respondents agreed that it did, which implies that it has face validity on the whole.

Table 8.7 Analysis results of the students' opinion of the face validity of the RS2

| | Frequency | Percentage | Cumulative percent |
|---|---|---|---|
| Strongly disagree | 0 | 0 | 0 |
| Slightly disagree | 7 | 8.9 | 8.9 |
| Slightly agree | 44 | 55.7 | 64.6 |
| Strongly agree | 28 | 35.4 | 100 |
| Total | 79 | 100.0 | |

I also analysed whether their answers correlated with their gender or their English score. The correlations between them were not significant, which means that there is no correlation between their conception of its face validity and either their gender or their English score.

Therefore, it can be concluded that over 90% of the respondents considered the RS2 to look valid on the whole, regardless of their gender or English score. I then analysed the question as to why they thought this. Those respondents who agreed with its face validity on the whole wrote that they did so because the scale appeared

[73] Since the questionnaire was not piloted in advance, this problem was not avoided.

to be systematic, clear, detailed and concrete (13),[74] because it assessed many aspects of writing (7), because it was an analytic scale which helped the students receive diagnostic feedback on their writing (5), and because it just looked great (3). On the other hand, those who disagreed with its face validity (around 10% of all the respondents) replied that it was because the RS2 included the aspect that raters would be likely to rate subjectively despite using the scale (3), because the terms such as 'often' and 'sometimes' were ambiguous and would result in ambiguity in band differentiation (3) and because the RS2 covered only the average level rather than all possible characteristics of each level (1). However, these negative views were expressed in less than 10% of all the answers to this question.

Second, I investigated whether they thought that the RS2 assessed appropriate features for writing assessment. As Table 8.8 below indicates, 92.4% of respondents agreed either slightly or strongly to this question.

Table 8.8 Analysis results of the students' opinion as to whether they thought that the RS2 assessed appropriate features

| | Frequency | Percentage | Cumulative percentage |
|---|---|---|---|
| Strongly disagree | 1 | 1.3 | 1.3 |
| Slightly disagree | 5 | 6.3 | 7.6 |
| Slightly agree | 41 | 51.9 | 59.5 |
| Strongly agree | 32 | 40.5 | 100.0 |
| Total | 79 | 100.0 | |

As in the first point, I analysed the correlations between their conceptions of the assessment features in the rating scale and either their gender or their English score, and they were not found statistically significant, either. Therefore, it is summarised that, regardless of their gender or English score, most respondents thought that the scale assessed features which were necessary and appropriate for writing assessment. In addition to this, I analysed the reasons for their response to this question. According to their answers, they thought that the rating scale included features which they found important, such as grammatical errors, sentence construction, connectors, the appropriateness of words used in a context, verbs, development, organisation and coherence (10). When the respondents who answered this question negatively were

[74] The figures in the brackets show the frequency of cases for each category.

asked to give the reason for their response, only one relevant answer was forthcoming. The rest either did not answer the follow-up question or gave irrelevant answers. The only useful response was that the rating scale did not deal with the appropriateness of register used in a context (1).

Furthermore, in Question 8 they were asked what features they thought were either inappropriate or needed to be added to. This question was originally supposed to mean inappropriate "assessment features or assessment categories" such as 'coherence', 'length' and 'development'. Unfortunately, however, they appeared to have understood it as inappropriate "descriptors", so their answers were not what I expected.[75] On the other hand, they did give some useful answers on aspects which need to be added to, such as 'appropriateness of word used in a context' (4) and 'native speaker-likeness of expressions used' (2).[76]

Given this, it is concluded that the RS2 is considered to assess appropriate features for writing assessment, including features such as 'coherence' and 'verbs', and that there are no assessment features which were found critically inappropriate.

Third, students were asked whether they thought that the demarcation between bands was done well. As the result in Table 8.9 below shows, over 84% of the answers to this question were positive.

---

[75] Even so, having a look at them for reference, they thought the following descriptors inappropriate: to give relatively low bands because a script is written in plain English (5); and to give a higher band between Bands 3 and 4 because it is long, even though a script has many errors (3). However, these two points seemed to result from the fact that they did not understand the term as intended, which may be because they were not provided with the manual for the RS2. For the former, "plain" was intended to mean "not appearing to be English used by native English speakers, but appearing to be English that is literally translated from Korean, even though it is not grammatically problematic" rather than "common" or "easy" which they may have taken it to mean. Even so, I admit that since five students pointed this out, the term needs to be replaced by a clearer term. For the latter, the difference between Bands 3 and 4 in Fluency not only relies on the difference in quantity but other aspects (see the RS2). Therefore, it seems that this point is caused by having insufficient time to fully understand the RS2.

[76] These were actually already included in the RS2, so it appears that they need to be better worded.

Table 8.9 Analysis results of the students' opinion of the differentiation between bands in the RS2

| | Frequency | Percentage | Cumulative percentage |
|---|---|---|---|
| Strongly disagree | 0 | 0 | 0 |
| Slightly disagree | 12 | 15.2 | 15.2 |
| Slightly agree | 44 | 55.7 | 70.9 |
| Strongly agree | 23 | 29.1 | 100 |
| Total | 79 | 100.0 | |

The correlations between their perception of band differentiation and either gender or English score did not turn out to be significant, either.

The reasons given for their positive answers were because the differentiation between bands looked clear (6); because there were neither too few nor too many bands (3); because the differentiation was based on various assessment aspects (1); and because the difference between bands looked reasonable (1).

The percentage of positive answers is still very high, as for the two points above. However, the percentage of negative answers on this point is slightly greater than that in the previous two points. The students who answered negatively gave the following reasons for their response: the differences between bands are not clear (9); the assessment is likely to be very subjective rather than objective and reliable, because the descriptors are not made up of concrete numbers (such as "one or two" and "five or six times") but ambiguous quantifiers or frequency adverbs (such as "sometimes" and "often") (8); there should be more than six bands to help differentiate between levels of writing ability (6); there are too many bands (4).[77] Thus, their greatest concern seemed to be the possibility that the assessment may be too subjective and unreliable because the descriptors are made up of unclear quantifiers and frequency adverbs, which do not indicate exact numbers.

Although I tried to exclude ambiguous words that could cause unreliability and subjectivity, when developing the RS2, this was not entirely possible. There could be two possible alternatives to using ambiguous frequency adverbs and quantifiers: differentiating band levels according to the difficulty of the task; and using exact numbers for observed occurrence. Unfortunately, neither appeared to be a proper

[77] As shown in Table 8.22, there were twelve negative answers to this question. The total number of reasons given here is greater than twelve because each respondent suggested more than two reasons for their response.

solution to this issue for the given situation in Korea. The first option did not seem appropriate for a rating scale that was to be used for a classroom testing situation, especially in Korea, where only one or at most two tasks are assigned to students in one assessment session, rather than the various types of tasks, as in general proficiency standard tests. Put another way, students are not asked to do enough different tasks to indicate the degree of difficulty of the tasks they can perform best. This kind of scale differentiating levels according the degree of difficulty of tasks could possibly be used at schools as well, but it would be for summative reports after students have been given and have performed various tasks during the course, rather than for each assessment session during the course, which is the concern of this study.

As for the second option, this did not seem appropriate, either. It was not possible to employ this choice: firstly, because there are some writing features, such as 'smooth connection between sentences', of which degree is not likely to be indicated by numbers; secondly, because even Korean students in the highest band made errors, as revealed by the statistical analysis done in Chapter Seven. The difficulty caused by these two reasons would not be surmountable, even if Upshur and Turner's (1995) EBB scale approach (see section 4.5.3) was adopted. Unambiguous dichotomous questions taking advantage of exact numbers or words, such as, "Is it coherent across the script?" and "Does it have no grammatical errors?" for the highest band cannot be developed as far as the descriptors were meant to be developed on the basis of their writing. Descriptors including words indicating degree, such as "Is it *fairly* good in terms of coherence across the script?" and "Does it have *few* grammatical errors?" would be more realistic.

In conclusion, these two alternatives were not appropriate for this situation. To deal with this issue, another method will be needed, which would require further research. Consequently, the use of frequency adverbs and quantifiers for this situation appeared unavoidable at the current stage, and the interpretation of these terms should be discussed beforehand by the developer, the users of the scale and the raters.

Finally, I investigated through Question 11 whether students thought that the RS2 would provide a valid picture of their writing ability. According to the statistical results presented in Table 8.10 below, many of them (84.8%) slightly or strongly

agreed that it would.

Table 8.10 Analysis results of the students' opinion as to whether the RS2 would provide a valid picture of their writing ability

| | Frequency | Percentage | Cumulative percentage |
|---|---|---|---|
| Strongly disagree | 0 | 0 | 0 |
| Slightly disagree | 12 | 15.2 | 15.2 |
| Slightly agree | 45 | 57.0 | 72.2 |
| Strongly agree | 22 | 27.8 | 100 |
| Total | 79 | 100.0 | |

As with the three previous points, there was no correlation between either their gender or their English scores and their opinion about the possibility of providing a valid picture of their writing ability.

In addition, it was found that respondents gave positive answers to Question 12 for the following reasons: the RS2 is an analytic scale which allows diagnostic feedback on writing (10); the RS2 looks systematic and clear (7); the ratings from self-assessment coincide with those from the RS2 (4); the assessment according to this RS2 is likely to be more reliable and objective than subjective judgement (2). On the other hand, the respondents who answered this question negatively did so for the following reasons: the raters' subjective judgement would inevitably come into play with the RS2 because of its ambiguous quantifiers and frequency adverbs (13); to assess a script *per se* is difficult and tricky, compared with scoring multiple-choice questions (6); the RS2 does not have all the necessary assessment aspects, for example register depending on genre (3); the rating scale only covers the average features of each level, whilst writing ability could have idiosyncratic characteristics across more than two levels at the same time (2); there are insufficient bands in the RS2 (1). While I do not disagree with these issues, many were hard to address due to the theoretical and practical limitations discussed above. Regarding the point that the RS2 cannot deal with idiosyncratic characteristics spanning two or more levels because it only covers the average features of each level, I think it is worth remembering that the aim of common rating schemes is fair judgement of writing skills, which means that they have to deal with the average, general characteristics of each proficiency level as a standard, and will never be able to cover every

idiosyncratic case. Attempting to do so would make them so unwieldy that they would be very hard to understand and use in a real testing context. As noted above, there will always be scripts that challenge even the best constructed rating scale, in which case raters will have to do their best to apply the rating scale as far and as well as they can.

One of the methods used to examine the validity of the RS2 was a questionnaire survey of students, who are one of main stakeholders in language testing. Regardless of their gender or English score, most answered that the RS2 appeared to have face validity in that it assessed features appropriate for writing assessment; that the band differentiation in the RS2 was clear and reasonable; and that if this RS2 was used, it would provide a valid picture of their writing ability. Therefore, I can conclude that the RS2 was conceived as valid according to the test-takers' judgement.

As discussed above, I used various methods to investigate the practicality, reliability and validity of the RS2. This investigation revealed that while there were some negative findings indicating that these three aspects needed further revision, there were also many positive findings. In the next section, which covers the final investigation into the RS2, I will discuss how it differs from FCE rating scale.

## 8.5 Comparison of the RS2 and the FCE rating scale

In Chapter One I discussed why there is a need to develop a rating scale when other published rating scales already exist. My arguments were that, first, every rating scale has its own specific context for which it is valid; second, published rating scales do not reflect the characteristics of English written by Korean students; and third, because the existing rating scales have defects as *a priori*-developed rating scales, as discussed in section 4.5.2. I will now consider whether the RS2 is different from the FCE scale and preferable in such respects.

First, the context and construct for which the FCE rating scale is valid are different from those for this RS2. The FCE scale is specifically designed for the FCE test, a kind of standardised test for an intermediate level, and the construct underlying the FCE rating scale is the ability to write accurately, appropriately and interestingly for target readers, commanding a range of vocabulary and expressions.

For this purpose, the FCE scale is constructed to have specific assessment categories and specific descriptors. However, the context and construct in the English Writing course are different from those of the FCE scale. The context in question is classroom assessment of a specific course, the English Writing course at FLHSK, and the construct implicit in the course is to write in an organised, accurate and fluent manner to express one's feelings and ideas across various genres. The RS2 was specifically developed, with these in mind, because scales should be context-specific, according to Hamp-Lyons (1991) and Storch (1993), as discussed in section 4.3, and Davies and Elder (2005) as discussed in section 3.3.1.2. Therefore, the two scales differ from each other (although certain features are common to the two constructs). Unlike the FCE scale, the RS2 includes the three writing features implicit in the course objectives as assessment categories in it for the sake of *a priori* construct validity, and given that it intends to be used in a classroom testing situation, it includes the descriptors which could provide Korean students with detailed and substantially helpful feedback, which were created from analysis of their writing, to help improve their English writing. A report/feedback could include a total band, three sub-bands and detailed feedback/comment for the three assessment categories. For example, a report/feedback for a student who is assigned Band 4 for Accuracy could be developed as shown below, taking advantage of the descriptors of the Band and the more detailed language of the coding scheme; and the same principle applied to Fluency and Organisation as well.

> For Accuracy, your English is generally good, but there are grammatical errors across various grammatical features in your writing. In particular, you need to be careful in the use of verbs. You sometimes either use more than one finite verb (e.g., *go, goes, went*) in a clause or omit it, and you appear to be confused between the usage of finite verbs and non-finite verbs (e.g., *going, gone, to go*) in a clause. In addition, you sometimes appear to be confused about verb types, such as intransitive verbs and transitive verbs, which causes you to omit or repeat the required complement(s) of a verb. Another observation is that there are many literal translations from Korean into English. These do not always make a clause/sentence ungrammatical, but they look awkward. You appear to need to be aware of English-like expressions. Clauses that are unintelligible due to serious errors with clause construction are also found in your writing. Therefore, it appears that you need to be careful about clause construction.

Second, the two rating scales differ from each other in terms of whether they

reflect the characteristics of both Korean students' English writing and Korean teachers' assessment. As for the former, I discussed it in section 1.6.4.3.2 as one of reasons why published rating scales are inappropriate to the context in Korea. For example, the diary study revealed that awkward expressions were found in Korean students' scripts. Such expressions were not ungrammatical, but created a negative impression, and therefore lead them to be awarded lower marks. However, the FCE scale includes neither concrete assessment categories nor descriptors to deal with this aspect. It only includes grammatical accuracy and the range of vocabulary/phrase used. The RS2, on the other hand, does include this feature. In Accuracy, there are descriptors regarding imperfection for even the top band, such as "Very accurate with few grammatical errors…" (Accuracy for Band 6); regarding awkwardness, such as "Good, but with a number of literal translations from Korean to English words… " (Accuracy for Band 4); and regarding idiosyncrasy, such as "Follows English word order, i.e. S+V+O/C rather than Korean word order, i.e. S+O/C+V, even though it is very poor with a great number of grammatical errors" (Accuracy for Band 2). The descriptors for Fluency also reflect the features of Korean students' writing by identifying specific aspects observed in high-level writing by Korean students, as in "Advanced level of language using advanced grammatical features (e.g., aspect, relative adverbs [rather than other features])..." (Fluency for Band 6) and "Good cohesion through little repetitions, use of correct pronouns and connectors beyond those at middle school level [rather than through other cohesive devices such as ellipsis and lexical cohesion]" (Fluency for Band 5). Organisation also consisted of features observed in Korean students' writing, with regard to content, format and so on, as in "No paragraphing and, if any, errors with paragraphing [rather than this feature is observed either in writing of other bands or not in any writing at all]" (Organisation for Band 3). These features are specific to Korean students, in that some of them cover the idiosyncratic features of writing by Korean students, while others point out features that are observed in a specific level of writing rather than in different levels of writing done by speakers of other languages. By aiming for an accurate reflection of Korean students' writing in this way, I tried to make the RS2 realistic and appropriate for use in the Korean context.

Additionally, as for the latter, I tried to make the RS2 specifically reflect Korean

teachers' characteristics in writing assessment. As discussed at the end of section 6.2.3.2, for example, they focused on length and/or grammar for the lowest level, and on advanced expression for the highest band; they hesitated between two candidate bands before deciding one of them; they did not find 'Register' in the FCE scale helpful for the assessment of their students' writing whilst they found that it was desirable to include a category regarding awkward English observed in their students' writing. These do not all seem to be reflected in the FCE scale. Thus, insofar as the RS2 reflects the characteristics of Korean students and teachers, it is different from the FCE rating scale.

Third, the RS2 was developed to avoid the problems with *a priori* developed rating scales. That is, the latter are based on the assumption of developmental sequence in writing, refer to the level of native speakers for the top band, do not reflect the learning theory of spirality and have a problem with validity, whilst this is not the case with the former because it was developed on the basis of data.

Finally, the RS2 differs from the FCE rating scale in that the differentiation between bands in the RS2 is made both qualitatively and quantitatively. As discussed in section 4.5.2, one of the problems with published rating scales is that the band differentiation in the scales depends on the use of ambiguous quantifiers and frequency adverbs. That is, bands are distinguished from each other quantitatively and with much ambiguity. Unfortunately, the RS2 also has ambiguous quantifiers, and so fails to overcome this problem, as noted by the raters in their diaries, think-alouds and responses to Questionnaire V, and by the students in Questionnaire VI. However, the bands in the RS2 are differentiated from each other qualitatively as well as quantitatively. That is, the distinction is not only by 'how much' a feature is observed but also by 'which feature' is observed in a band. Therefore, the RS2 includes 'which feature is observed' for a specific band as well as 'how much/often' a feature is observed for a band. These results from the development principle behind this RS2, that its descriptors are based on features discriminating between neighbouring bands. Thus, for example, 'Disconnected and incoherent sentences for more than 50% of the whole script' in Fluency is not mentioned at all for Bands 6 and 5 because statistical analysis of the scripts in these two bands found that it was a non-distinctive feature between them, whereas it is mentioned as a distinctive feature

differentiating between Bands 2 and 3. This characteristic of the RS2 appears to accommodate the behaviour observed in both the findings from the raters' diary and the findings of Pollitt and Murray (1996) discussed in section 4.4, that raters pay attention to different features depending on the level of the test-taker's performance. Therefore, the RS2 appears to be appropriate as an assessor-oriented scale as intended.

Thus, having compared the RS2 with the FCE rating scale, it can be seen that the two rating scales can be distinguished from each other in terms of the three aspects mentioned above.

## 8.6 Summary and conclusions

In this chapter I have dealt with the empirical investigation of the RS2 in terms of practicality, reliability and validity, and discussed the differences between the RS2 and the FCE rating scale. To begin with, the differences between these two scales are as follows: the construct and context for which the FCE rating scale is valid are different from those for the RS2; the RS2 reflects the characteristics of Korean students' writing and Korean teachers' concern and behaviour in writing assessments, unlike the FCE scale; most of the problems with *a priori*-developed scales have been removed from the RS2; and the band differentiation in the RS2 is both qualitative and quantitative, whereas the FCE rating scale depends on mainly quantitative aspects.

The validation of the RS2 was investigated in terms of three aspects: practicality, reliability and validity. As with Harrison (1969), three aspects were investigated with regard to practicality: economy, ease of interpretation and ease of administration and scoring. For the enquiry into the practicality of the RS2, I developed a questionnaire and asked the three raters to respond to it. The analysis of their answers to the questionnaire revealed that the raters found the RS2 practical on the whole, although there gave some negative feedback as well.

The investigation of reliability was dealt with in the realm of validity, with the investigation focusing on construct validity, which tends to be considered as a central overarching concept. The construct validity of the RS2 was investigated both quantitatively and qualitatively in terms of seven methods: a correlation study, a

MTMM study, a "G-study" (ANOVA), two questionnaire studies, a diary study and a think-aloud study. In the correlation study, intra-rater reliability was calculated and found to be high, higher than in the cases of subjective holistic assessment and the FCE rating.

In the MTMM study, it was found that the RS2 had both convergent and discriminant evidence of construct validity. Meanwhile, there were high correlations between the ratings given using the RS2 and the FCE rating scale. This was presumably because both scales assess writing skill, even though their constructs of writing abilities are different from each other and consequently include different assessment categories. The fact that the RS2 has concurrent criterion-related evidence of construct validity with the FCE rating scale might result in the conclusion that there is no need to develop the RS2. However, with regard to the construct validation of the RS2, it also means that the RS2 has construct validity as a rating scale for writing assessment, given that it has concurrent criterion-related evidence of construct validity with the FCE rating scale, which is a recognised rating scale for writing assessment. Neutrally stated, since this high correlation could result from a hidden variable, this does not mean that the two scales measure the same things, as Fulcher (2003) points out about his own similar findings.

The validity of the RS2 was also checked by qualitative methods. A questionnaire study conducted on the three raters found that they believed (1) that they understood the scale properly as intended with the aid of the manual; (2) that the scale helped them be consistent in considering the appropriate categories for writing assessment; (3) that the three main categories in the scale appeared to be assessed properly; and (4) that the scale appeared to provide a valid picture of Korean learners' writing ability. However, there was negative feedback on the scale as well, which needs to be considered in a further study. For example, Rater A said that he did not think that fluency was assessed properly because he found the features in Fluency ambiguous. The answers to the questionnaire indicate that, with the exception of a few points, the raters generally understood and applied the scale as intended; that they thought the scale dealt with the three main assessment categories properly, apart from Fluency which requires subjective judgements (with Rater A); and that their ratings generally appeared to be valid.

These findings were further investigated using methods that helped reveal the rating process followed by each rater – a diary study and a think-aloud study. Of their answers to Questionnaire V, Rater A said that they thought they understood the scale except for the aspect of Fluency. These diary study and think-aloud study also revealed that their perception of Fluency was simpler than suggested. They found the quantifier, frequency adverbs and other features in Fluency, apart from 'length' and 'language level' ambiguous and hard to understand, although Accuracy and Organisation were easy to deal with on the whole. These studies also showed that the raters sometimes understood certain features either differently from or more simply than the way that was suggested, and that they were sometimes influenced by their own subjective criteria.

Nevertheless, since all of the raters understood the essentials of the scale, the inter-rater reliability coefficients appeared to be high. In addition, given the results of "G-study" (ANOVA), the facet of rater did not affect the score and its construct validity.

All of these validation methods were those carried out with the raters, who are teachers. Since learners are also stakeholders in the testing situation, it was important that they were included in the study too, so I conducted a questionnaire study with Korean students. Most responded that the scale appeared to be suitable for writing assessment and that it properly assessed the main assessment aspects.

In conclusion, the investigation revealed that both teachers and learners believed the RS2 to have many positive aspects towards construct validity, despite some negative features that need to be improved.

# CHAPTER NINE. SUMMARY AND CONCLUSIONS

## 9.1 Introduction

In Chapters One to Eight, I have reviewed the literature on writing assessment, rating scales and writing ability and presented the procedure and results of this study. In this chapter I will conclude this study. In section 9.2 I will summarise the main points and in section 9.3 I will discuss the limitations of this study and make suggestions for further research and development in this area.

## 9.2 Summary

As the need for writing skills became more and more apparent, an English Writing course for high schools in Korea was established for the first time in 1997 as part of the 7th national curriculum. As a first step, the course was only introduced to FLHSK. However, although the course was established and textbooks authorised by the Korean Ministry of Education were published, there was no suggested rating scale to be used for the assessment of the course.

The assessment procedure which the guidelines to the course suggest is to assess according to the teachers' own subjective holistic criteria, rather than suggesting a new rating scale or one of the published rating scales. However, this is not in fact a recommendable method because it results in problems with reliability and validity.

Published rating scales are not appropriate for this context, either, because their constructs and the purposes and contexts of assessment are not exactly coincident with those of the assessment of the course. The scales also do not reflect the characteristics of Korean students and teachers, but are for L2 learners worldwide. In addition, they have drawbacks as *a priori*-developed rating scales, in particular for example, the descriptors in the scale are not based on an empirically derived model.

This study aimed, therefore, to develop a rating scale for writing assessment for the course of English Writing at FLHSK. When it comes to the approach to the rating scheme development, considering the problems of rating scales developed *a priori*, this new rating scale was to be developed by a data-based approach.

To begin with, I undertook a questionnaire survey of one hundred and four

English teachers at fourteen FLHSK to find out whether they did writing assessment in their class and if they thought a rating scale needed to be developed specifically for Korean students. With generally positive answers to the questionnaire, I began the development of the rating scale.

First, before trying to obtain writing samples for a data-based approach, I did a pilot study to determine appropriate writing tasks for the students. I prepared two kinds of tasks, an informal letter and a formal essay. The topics for these tasks were devised on the basis of the textbooks for this course. After asking three hundred and thirty-three students at Kwacheon foreign language high school to do these tasks, I tried to find out through a questionnaire survey how they found the tasks. The tasks were found to be appropriate, but some points in the prompts needed to be revised.

Second, with these findings, I prepared the same kind of tasks with different topics, revising some points in the prompts. The same students as in the pilot study did two kinds of writing tasks over two sessions. They were informed of the topics of the tasks beforehand. As a result, I obtained six hundred and sixteen scripts from them. Unfortunately, however, since I could not find more than three raters to rate the scripts for this study, only three hundred and ninety of them, which the three raters, Raters A, B and C managed to rate, were used for this study.

Third, I asked the three raters to rate the scripts (Groups A, B and C) according to their own subjective criteria. They were instructed to keep a diary and do a think-aloud for their rating process. In addition, I asked them to rate some of the scripts (Groups D, E, F and G) according to their subjective criteria for the investigation of both inter-and intra-rater reliabilities. In the same way, I asked them to use the FCE rating scale. The FCE rating scale was chosen among published rating scales because it seemed the most likely candidate for this situation in that it was for intermediate level.

These procedures were to empirically investigate the tentative assumption made in section 1.2 that subjective holistic scoring had problems in reliability and validity and that the FCE rating scale was not appropriate for the Korean situation. From these procedures, I found that whilst both intra-and inter-rater reliabilities in both subjective holistic scoring and the FCE rating scale were fairly high, except for the intra-rater reliability of Rater A, the certain weakness of these two scoring methods

were revealed by the diary and verbal protocols. When doing subjective holistic scoring, the raters considered features irrelevant to the object of the course, were not sure of their rating, had somewhat vague criteria for the highest and the lowest bands and were affected by the scoring of previous scripts. When using the FCE rating scale, on the other hand, the raters brought in their own criteria. Since raters are unlikely to put aside their own subjective value completely even when they are using a rating scale, this phenomenon is almost inevitable, but the real problem was that they were not familiar with certain concepts such as Appropriacy of Register, Range, Target reader and Organisation in the FCE scale, and understood them in a different way from intended, which could have affected the validity of their rating. They also felt that some assessment categories were unnecessary (Register) or needed to be added explicitly ('length', 'development' and 'awkward /literally translated expression from Korean'). They added that band differentiation mainly using quantifiers could cause inconsistency in interpretation and that the descriptors were appropriate only for the FCE writing test and, though logical, did not reflect some of the features which are often observed in Korean students' scripts. For these reasons, it was concluded that the FCE rating scale was not appropriate for the context in Korea.

Fourth, I therefore began to develop a rating scale for the context in question. I started by grouping the three hundred and ninety scripts into six groups according to the raters' ratings from subjective holistic scoring. To develop a coding scheme for the scripts, I repeatedly read them in groups, took into consideration the construct of writing ability implicit in the objectives of the course and previous literature on writing and developed coding categories by which to code them. The coding scheme had Accuracy, Fluency and Organisation as its main categories, with eighty-two subcategories. After coding every script according to this coding scheme, I statistically analysed the coding to find the distinctive features of each pair of neighbouring bands. On the basis of these features, I constructed the RS1 which was an analytic scale of six bands of 1-to-6 including three main assessment categories (Accuracy, Fluency and Organisation). The scale was used by the three raters before being adjusted with some revisions according to their feedback through diary, verbal protocols and a questionnaire.

Fifth, the RS2 was used by the raters again, to examine its practicality, reliability and validity. For the investigation of practicality, I asked them to answer a questionnaire which I constructed on the basis of Harris (1969), to investigate economy, ease of interpretation and ease of administration and scoring. It was reported that the time to grasp the scale varied from thirty minutes to a few hours depending on the raters. Although Rater C made some negative comments, on the whole all of the raters, especially Raters A and B, found it understandable and practical to administer, except in some minor defaults.

For validity, in line with of the current trend to regard construct validity as a central unitary concept, with multiple types of evidence employed for validation, the construct validity of the RS2 was investigated through seven validation methods. These included both quantitative methods--correlational study, an MTMM and a "G-study" (ANOVA) -- and qualitative methods--questionnaires, diary and think-aloud.

Summarising the findings from these studies, the correlation coefficients of intra-rater reliability was high, and this was the case with even Rater A whose intra-rater reliability for both subjective holistic scoring and the FCE rating scale scoring was not significant. This led to the conclusion that the RS2 may have helped the raters sustain internal consistency while using it.

The MTMM showed the evidence for both convergent and discriminant evidence of validity of the RS2, when it was judged according to four criteria for multitrait-multimethod judgements that was set up by Campbell and Fiske (1959).

According to their questionnaires, the raters found the RS2 appropriate for the given construct, and thought that it could provide a valid picture of Korean students' writing ability and that they had understood the scale as intended on the whole, even though they had difficulty applying the concept of Fluency.

According to the students' questionnaires, they also liked the RS2. Regardless of their gender and English score, most of them found the scale reasonable, with systematic, clear, concrete, varied, analytic and appropriate assessment features. Therefore, most of them thought that the RS2 could provide valid assessment of their writing ability.

The validity of the RS2 was examined by looking at the rating process again. For this, a diary study and a verbal protocol study were carried out. It was found that

the raters understood the RS2 as intended, by and large. They were sometimes affected by their subjective criteria in the way of introducing and applying assessment categories in their own criteria when using the RS2. However, this was not unique to this case, and seems likely to happen whenever a rating scale is used, given that it was also observed when they were using the FCE rating scale and it is supported by previous study. Also, the raters did not mention all of the features within each main assessment category. Although they focused on certain parts of each assessment category, especially in the case of Fluency, those features which they did mention suggest that they understood the essentials of the category, which provide convergent evidence for validity of this scale, given that the raters were not trained other than self-training through manual. In addition, except for Fluency and particularly 'coherence' in Fluency, it did not appear that they had difficulty understanding the concepts of the features. However, they found the meaning of quantifiers and frequency adverbs ambiguous and unclear. This problem has not been solved in other rating scales, as mentioned in section 4.5.2, which would be an area for which further research need to be done.

Of course all outcomes of the rating process slightly varied depending on the raters. Nonetheless, according to the "G-study" (ANOVA), the facet of rater did not greatly affect the score. Put another way, the difference amongst the raters in interpreting and applying the RS2 was not significant.

I conclude from the above that the RS2 does have practicality and validity on the whole and that it can be used in preference to the FCE scale. The most important difference between them is that they are valid for different contexts and constructs. The given context is for a kind of course-based classroom testing in the course of English Writing at high school level in Korea, with the construct being the ability to write various genres of writing in an organised, accurate and fluent manner. On the other hand, the context for the FCE rating scale is for the FCE writing assessment which is a kind of proficiency test developed for L2 learners worldwide at intermediate level, with the construct being the ability to write given genres of writing accurately, appropriately and interestingly to target readers, commanding a range of vocabulary and expressions. Just as the FCE rating scale was developed for this specific context to do valid assessment even though other published rating scales

existed, the RS2 which is valid for the context in question fulfils a need.

Furthermore, the RS2 was different from the FCE rating scale in that the former reflected the characteristics of Korean students' English writing and Korean teachers' rating behaviour whilst the latter was not specifically developed so.

In addition, the RS2 was developed through a data-based approach, which helped reflect the features of real writing by Korean students rather than making assumptions according to the logic of writing development, and also helped avoid the problems associated with the *a priori*-developed rating scales.

There was another aspect in which the RS2 was different from the FCE rating scale. Whilst the latter differentiated between bands in terms of quantitative aspects, the former did so in terms of qualitative as well as quantitative aspects. That is, the band differentiation in the latter had to do with "how many/much" a given feature there are for a band using ambiguous quantifiers. On the other hand, the band differences in the former had to do with "what feature" as well as "how many/much" a feature there are for a band.

## 9.3 Limitations of the study and suggestions for further development

Inevitably, in spite of my best efforts while conducting the study, some limitations exist. I will now discuss eight of these and also make suggestions for further development of the scale in the future, so that it can be used in other schools in Korea and to make it more teacher/user-friendly.

First, when obtaining the scripts which were analysed for rating scale development, I decided on two kinds of tasks: an informal letter and a formal essay. They were chosen given that they were considered to be to some extent representative and common genres and tasks for either informal or formal continuous writing. Even so, if the scripts had been from more various tasks and genres than these two, the features of Korean students' writing may have been different from those observed in this study. In order to make the scale more generalisable, there is a need to use a wide range of genres and tasks.

Second, the students who did the writing tasks for this study were informed of the topics of the tasks beforehand. Although I had intended not to do so because of the concern that they might be helped by others or by looking at materials rather than

preparing by themselves, due to the custom and practical limitations in the classes, according to the suggestions of teachers who were in charge of the classes, the students were informed of them. As a result, the students were allowed to prepare for their writing at home. Fortunately, however, the teachers confirmed that the students had prepared by themselves under similar circumstances. Nonetheless, since performance in a test situation where they are not informed of topics in advance may be different from that in a non-test situation where they are allowed to prepare for a writing task with the notification of the prompt and since the RS2 may be used for test situations, there is a need to obtain scripts without students being notified beforehand.

Third, the scripts for this study were obtained from students at only one foreign language high school, of more than twenty in Korea, for practical reasons that I was not able to access students at other schools. For the scale to be more generalisable, the scripts which form the basis of the development of the rating scale will need to be obtained from various schools.

Fourth, only three English teachers took part in the study as raters. They rated the obtained scripts according to their own subjective criteria, using the FCE rating scale and using the RS2. As a result, the feedback and the information on rating behaviour obtained from them were limited, although the raters varied in terms of career histories and schools where they worked. More raters will need to take part in future studies to provide more extensive information that would help improve the scale. Furthermore, only one of the raters for this study worked at a foreign language high school, whilst the others worked at general academic high schools. As the RS2 was developed for the current course at foreign language high schools, I tried to get all the raters for this study from foreign language high schools, but this was not possible due to the difficulty of finding sufficient raters for this study. I would not say, however, that this variable greatly affected the entire study. The two raters from general academic high schools may have adjusted their expectations of the level of students' writing while they were doing the pilot assessments prior to the main assessments of each scoring. Nevertheless, if all the raters had worked at foreign language high schools, as I would have preferred, their feedback could have been considered to be more directly relevant to the judgement of the quality of the RS2.

Fifth, the factor, 'task' was not included in the "G-study" (ANOVA) and the MTMM that were used as quantitative validation methods for the RS2. Given that 'task' is one of the major variables in writing assessment, along with the rater variable for these two studies, the research plan should have been established more carefully.

Sixth, with regard to blind coding as a method for quality control in data coding, I employed twenty scripts chosen at random. This was because I thought random sampling would be one way to avoid the effect of the variable of script level affecting quality control in data coding. However, it might have been better to employ specifically chosen scripts, such as Group D, for which the raters showed quite high inter-rater reliability in their subjective holistic scoring. It would be meaningful to investigate the agreement between the coders, compared to the high agreement in the raters' assessment.

Seventh, with regard to the use of quantifiers, these have been found in other studies not to be helpful for the raters' proper understanding and use of rating scales as intended since they do not indicate definite number or quantity. It was the same case in this study, and whilst I made distinctions between bands qualitatively, since I also did quantitatively using the quantifiers, the raters found them ambiguous. More research needs to be done on this problem.

Finally, to improve raters' application of the scale, especially Fluency needs to be revised to be clearer. The concept of fluency is more abstract than, for example accuracy on its own. Consequently, to operationalise and analyse the former was more difficult than the latter, so that this difficulty has been discussed in previous studies as well. For this study, this problem was also detected at the stage of both coding scripts and quality control in coding data as discussed in Chapter Seven. It seems that raters also had similar difficulty with Fluency in the RS2. In order to make Fluency clearer and easier to use, more research on fluency, especially the operationalisation of the concept, needs to be done.

Of the eight points mentioned above, the last two points regarding the use of quantifiers and the operationalisation of fluency need deeper research to resolve. On the other hand, the first six points were caused by practical limitations of this study. Therefore, if research is carried out more extensively, cooperatively and carefully, it

is likely that these issues will be able to be solved.

## 9.4 Implications and suggestions for further research and development

Space permits mention of only a few of the many possibilities for further research. One would be to investigate the relationship between rater's background, and their rating behaviour and views on good writing, and to find out if there are any features of the scale which are not easily accepted by raters and may need to be adjusted, depending on their background. In the present research, the three raters had varied backgrounds: all were English teachers at high schools, but they differed in some respects such as the teaching career history and school types which they worked for. These differences may have had an effect on their rating behaviour and views on good writing. When trying to develop a new rating scale or revise the RS2, therefore, this would be one of the facets to be investigated. Furthermore, whilst only Korean English teachers participated in the study, given that the writing course may be taught by English native speakers, it is also desirable to do research on these in order to find out their view on good writing, to obtain their feedback on the scale in question and to compare it to that obtained from Korean raters. This consideration of rater background in rating scale development could be undertaken into in other countries as well, especially when they try to develop a rating scale specific to their own circumstances.

Second, there is a need to investigate if there is any difference between what Korean English teachers theoretically conceive as good writing and what they really pay attention to while rating, and to find out if their view on good writing is affected by practical constraints, for example having little time to assess writing, or having large classes to do. Since such practical constraints, if any, are unlikely to change in a short period, this kind of study would try to find solutions for this, in developing a rating scale.

Third, it would be desirable, when devising writing tasks and reporting ratings to take account students' opinions in order to increase their interest and motive and to help improve their writing ability. For the tasks, according to the students' feedback in this study (see section 5.5.1), they found it useful to have key points for content which could help them spend less time in deciding what to write about for a given

topic. For the topic of the task, I chose from textbooks for the English Writing course (since this study had to do with assessment in the course, I accepted the textbooks as they were without any criticism on it). According to the teachers who helped collect the writing samples, unfortunately, the students did not find the topics interesting enough to motivate them. This suggests that there is a need that textbook developers consider students' interests to help motivate them and to allow them to have many things to write about when choosing topics in the textbooks. In addition, as for reporting method, it was found from the questionnaire study to the students that they wanted to receive a detailed feedback, to help them improve their writing ability by becoming aware of their weaknesses and strengths in detail (see section 5.5.1). Teachers, therefore, can be advised to meet this need as much as possible when assigning ratings for the effective teaching and learning of writing.

Fourth, writing assessment is not included in university entrance examinations in Korea, which have a very significant backwash effect on education from primary to high-school level. Therefore, before aiming at any backwash effect on the teaching of writing by including it in the exam, a rating procedure would need to be made easy to use. Given that writing assessment generally requires more time, effort and human resources than the assessments for other skills, this issue of feasibility would require careful exploration. For nationwide examinations as well as for classroom testing, which was the concern of this study, there is a need to develop a rating scheme in which this is taken into consideration. For this, it would perhaps be helpful to consider/compare rating scales and the situation in other Asian counties such as Japan[78] in which education is also exam-oriented.

Fifth, to encourage the teaching and assessing of writing, a long-term plan for teacher support needs to be made by the central authorities in Korea. Because writing assessment tends to require relatively more time, effort and resources to administer than assessment for other skills, it has not so far been done very much. I believe, however, that considering the current situation that many Korean people need to be able to write in English fluently and effectively since they have a great number of opportunities to go abroad for study and there are a great deal of correspondence in

[78] Japan is an EFL context like Korea and writing assessment is carried out for private university entrance examinations only, not for national university entrance examination.

written form in international communities, the Ministry of Education of Korea needs to make efforts, in addition to have the recently introduced English Writing course, to actively support the teaching and assessing of writing in the public education sector. To this end, the development of a national rating scale should be one of the first steps, and in-service education and concrete guidelines on how to teach and assess students' English writing need to be provided.

Sixth and finally, many studies on writing, including this one, have been carried out cross-sectionally. Fulcher (1997) notes White's (1989) longitudinal study with L1 children, and argues that longitudinal studies also need to be done with L2 writers. As he suggests, conducting such studies with L2 learners of various ages and backgrounds in many parts of the world would contribute to our understanding of the acquisition of writing skills and help improve the assessment of writing.

We can expect, in the next few years, even more research, development and debate on the assessment of writing—on an international level, emphasising theory-building and generalisability, and on national and local level, emphasising practicality, acceptability and washback. The present thesis cannot, of course, offer final answers, but I hope it will contribute to the debate.

# BIBLIOGRAPHY


**In Korean**

Lee, B.-h., Choi, J.-h., Park, K.-h. and Kim, E.-j. (2001). "Ko-deung-hak-kyo Kyo-yook-kwa-cheong Hae-seol 11: oi-kook-eo (young-eo)" [*Guideline to curriulum at high schools 11:Foreign language (English)*]. Seoul, Korea: Ministry of Education of Korea.

**In English**

Ackerman, J. M. (1990). Students' self-analyses and judges' perceptions: where do they agree? In L. Flower, V. Stein and *et al.* (Eds.), *Reading to Write: Exploring a cognitive and social process* (pp. 96-115). New York: Oxford University Press.

ACTFL. (1987). ACTFL proficiency guidelines. In H. Byrnes and M. Canale (Eds.), *Defining the Developing Proficiency: Guidelines, implementations and concepts* (pp. 5-24). Lincolnwood, IL: National Textbook Company.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson and B. North (Eds.), *Language Testing in the 1990s: The communicative legacy* (pp. 71-86). London: Macmillan.

Alderson, J. C. and Clapham, C. (1995). Assessing student performance in the ESL classroom. *TESOL Quarterly*, 29(1), 184-187.

Allen, P., Cummins, J., Mougeon, R. and Swain, M. (1983). *Development of Bilingual Proficiency: Second year report*. Toronto, Ont.: The Ontario Institute for Studies in Education.

Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York: Macmillan.

Angoff, W. H. (1988). Validity: an evolving concept. In H. Wainer and H. I. Braun (Eds.), *Test Validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Asher, A. L. (1983). *Language Acquisition diaries: Developing an awareness of personal learning strategies*. Unpublished master's thesis, School for International Training, Brattleboro, Vermont.

Bachman, L. F. (1985). Performance on cloze test scores. *TESOL Quarterly*, 16(1), 61-70.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bailey, K. M. (1990). The use of diary studies in teacher education programs. In J. C. Richards and D. Nunan (Eds.), *Second Language Teacher Education* (pp. 215-226). Cambridge: Cambridge University Press.

Bailey, K. M. and Ochsner, R. (1983). A methodological review of the diary studies: windmill tilting or social science? In K. M. Bailey, M. H. Long and S. Peck (Eds.), *Second Language Acquisition Studies* (pp. 188-198). Rowley, Mass.: Newbury House.

Bailey, N. (1989). Discourse conditioned tense variation. In M. R. Eisenstein (Ed.), *The Dynamic Interlanguage: Empirical studies in second language variation* (pp. 279-296). New York: Plenum.

Bardovi-Harlig, K. (2000). *Tense and Aspect in Second Language Acquisition: Form, meaning, and use*. Oxford: Blackwell.

Benson, C. (2002). Transfer/cross-linguistic influence. *ELT Journal,* 56(1), 68-70.

Bereiter, C. and Scardamalia, M. (1987). *The Psychology of Written Composition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Berlin, J. A. (1988). Rhetoric and ideology in the writing class. *College English*, 50, 477-494.

Bhardwaj, M., Dietrich, R. and Noyau, C. (1988). *Temporality* (Final Report to the European Science Foundation, Vol. V). Nijmegen: Max Planck Institute.

Black, L., Daiker, D. A., Sommers, J. and Stygall, G. (1992). *Handbook of Writing Portfolio Assessment: A program for college placement*. Oxford, OH: Department of English.

Briere, E. (1966). Quantity before quality in second language composition. *Language Learning*, 16, 141-151.

Brindley, G. (1998). Describing language development?: rating scales and SLA. In L. F. Bachman and A. D. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research* (pp. 112-140). Cambridge: Cambridge University Press.

Brossell, G. (1986). Current research and unanswered questions in writing assessment. In K. L. Greenberg, H. S. Wiener and R. A. Donovan (Eds.), *Writing Assessment: Issues and strategies* (pp. 168-182). New York: Longman.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12(1), 1-15.

Brown, H. D. (1987). *Principles of Language Learning and Teaching* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall Regents.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently?

*TESOL Quarterly*, 25(4), 587-603.

Brown, J. D., Hilgers, T. and Marsella, J. (1991). Essay prompts and topics: minimising the effect of mean differences. *Written Communication*, 8(4), 533-556.

Brumfit, C. J. (1984). *Communicative Methodology in Language Teaching: The roles of fluency and accuracy*. Cambridge: Cambridge University Press.

CAE Handbook. (2001). Cambridge: University of Cambridge ESOL Examinations.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Jr. Oller (Ed.), *Issues in Language Testing Research* (pp. 333-342). Rowley, Mass.: Newbury House.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.

Carlson, S. and Bridgeman, B. (1986). Testing ESL student writers. In K. L. Greenberg, H. S. Wiener and R. A. Donovan (Eds.), *Writing Assessment: Issues and strategies* (pp. 126-152). New York: Longman.

Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing*, 2(2), 153-190.

Carroll, B. J. (1980). *Testing Communicative Performance: An interim study*. Oxford: Pergamon Institute of English.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, 14, 3-22.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research* (pp. 32-70). Cambridge: Cambridge University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English*, 8(1), 65-81.

Choi, Y-H. (1988). Text structure of Korean speakers' argumentative essays in English. *World Englishes*, 7(2), 129-142.

Choi, Y.-H. (2000). Effects of writing test tasks on learner performance and rating. *English Teaching*, 55(3), 217-245.

Clark, J. L. (1985). Curriculum renewal in second-language learning: an overview. *The Canadian Modern Language Review*, 42(2), 342-360.

Clark, J. L. and Clifford, R. T. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques. *Studies in Second Language Acquisition*, 10, 129-147.

Cohen, A. (1984). On taking language tests: what the students report. *Language Testing*, 1, 70-81.

Cohen, L., Manion, L. and Morrison, K. (2000). *Research Methods in Education* (5th ed.). London: RoutledgeFalmer.

Collins Cobuild English Grammar. (1990). London: HarperCollins.

Conklin, E. L. (1982). *Writing Answers to Essay Questions: A naturalistic study of the writing process*. Unpublished doctoral dissertation, Indiana University of Pennsylvania.

Conlan, G. (1986). "Objective" measures of writing ability. In K. L. Greenberg, H. S. Wiener and R. A. Donovan (Eds.), *Writing Assessment: Issues and strategies* (pp. 109-125). New York: Longman.

Connor-Linton, J. (1995). Looking behind the curtain: what do L2 composition ratings really mean? *TESOL Quarterly*, 29(4), 762-765.

Converse, J. M. and Presser, S. (1986). *Survey Questions: Handcrafting the standardised questionnaire*. Thousand Oaks, Calif.: SAGE Publications.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

CPE Handbook. (2002). Cambridge: University of Cambridge ESOL Examinations.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-51.

Davies, A. (1995). Introduction: measures and reports. *Melbourne Papers in Language Testing*, 4(2), 1-11.

Davies, A. (1997). The competing claims of accuracy and fluency in the construction of performance tests of language proficiency: two cheers for Robert Lado! *Melbourne Papers in Language Testing*, 6(2), 1-19.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamamra, T. F. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Davies, A. and Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 795-813). Mahwah, NJ: Lawrence Erlbaum Associates.

DeRemer, M. L. (1998). Writing assessment: raters' elaboration of the rating task. *Assessing Writing*, 5(1), 7-29.

Downing, A. and Locke, P. (1992). *A University Course in English Grammar*. Hemel Hempstead: Prentice Hall.

Eggington, W. G. (1987). Written academic discourse in Korean: implications for effective communication. In U. Connor and R. B. Kaplan (Eds.), *Writing Across Languages: Analysis of L2 Texts* (pp. 153-168). Reading, MA: Addison-Wesley.

Emig, J. (1971). *The Composing Processes of Twelfth Graders*. Urbana, IL: National Council of Teachers of English.

Emig, J. (1983). *The Web of Meaning*. Upper Montclair, NJ: Boynton/Cook.

Ericsson, K. A. and Simon, H. A. (1993). *Protocol Analysis: Verbal reports as data* (2nd ed.). Cambridge, Mass.: The MIT Press.

Faigley, L., Daly, J. and Witte, S. (1981). The role of writing apprehension in writing performance and competence. *Journal of Educational Research*, 75, 16-21.

FCE Handbook. (2001). Cambridge: University of Cambridge ESOL Examinations.

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler and W. S.-Y. Wang (Eds.), *Individual Differences in Language Ability and Language Behaviour* (pp. 85-101). New York: Academic Press.

Flower, L. (1989). Cognition, context and theory building. *College Composition and Communication*, 40: 282-311.

Freedman, S. W. (1981). Influences on evaluators of expository essays: beyond the test. *Research in the Teaching of English*, 15, 244-255.

Fulcher, G. (1987). Tests of oral performance: the need for data-based criteria. *ELT Journal*, 41(4), 287-291.

Fulcher, G. (1993). *The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language*. Unpublished doctoral dissertation, University of Lancaster.

Fulcher, G. (1996a). Invalidating validity claims for the ACTFL Oral Rating Scale. *System*, 24(2), 163-172.

Fulcher, G. (1996b). Does thick description lead to smart tests?: a data-based approach to rating scale construction. *Language Testing*, 13(2), 208-238.

Fulcher, G. (1996c). Testing tasks: issues in task design and the group oral. *Language Testing*, 13(1), 23-51.

Fulcher, G. (1996d). Writing in the classroom. *Modern English Teacher*, 5(3), 45-48.

Fulcher, G. (1997). Assessing writing. In G. Fulcher (Ed.), *Writing in the English Language Classroom* (pp. 91-105). Hemel Hempstead: Phoenix ELT/Prentice Hall Macmillan.

Fulcher, G. (1999). Assessment in English for Academic Purposes: putting content validity in its place. *Applied Linguistics*, 20(2), 221-236.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Longman/Pearson

Education.

Galloway, V. B. (1987). From defining to developing proficiency: a look at the decisions. In H. Byrnes and M. Canale (Eds.), *Defining the Developing Proficiency: guidelines, implementations and concepts* (pp. 25-73). Lincolnwood, IL: National Textbook Company.

Gathercole, V. C. (1986). The acquisition of the present perfect: explaining differences in the speech of Scottish and American children. *Journal of Child Language*, 13, 537-560.

Gee, S. (1997). Teaching writing: a genre-based approach. In G. Fulcher (Ed.), *Writing in the English Language Classroom* (pp. 24-40). Hemel Hempstead: Prentice Hall Europe ELT.

Gillham, B. (2000). *Developing a Questionnaire*. London: Continuum.

Glaser, R. (1984). Education and thinking: the role of knowledge. *American Psychologist*, 39(2), 93-103.

Glendinning, E. and Howard, R. (2001). Examining the intangible process: Lotus ScreenCam as an aid to investigating student writing. *Edinburgh Working Papers in Applied Linguistics*, 11, 42-58.

Goetz, J. P. and LeCompte, M. D. (1984). *Ethnography and Qualitative Design in Educational Research*. New York: Academic Press.

Gowen, S. (1984). Writing, rating and personality type. Paper presented at the ninth annual University System of Georgia Developmental Studies Conference, Athens.

Grabe, W. and Kaplan, R. B. (1996). *Theory and Practice of Writing: An applied linguistic perspective*. Essex: Longman.

Green, A. (1998). *Verbal Protocol Analysis in Language Testing Research: A handbook* (Vol. 5). Cambridge: Cambridge University Press.

Greenberg, K. L. (1992). Validity and reliability issues in direct assessment of writing. *WPA: Writing program administration*, 16(1-2), 7-22.

Greenhalgh, C. and Townsend, D. (1981). Evaluating students' writing holistically---an alternative approach. *Language Arts*, 58(7), 811-822.

Griffin, P. E. (1990). Profiling literacy development: monitoring the accumulation of reading skills. *Australian Journal of Education*, 34(3), 290-311.

Haegeman, L. (1994). *Introduction to Government and Binding Theory* (2nd ed.). Oxford: Blackwell.

Hairston, M. (1982). The winds of change: Thomas Kuhn and the revolution in the teaching of writing. *College Composition and Communication*, 33(1), 76-88.

Hamilton, J., Lopes, M., McNamara, T. and Sheridan, E. (1993). Rating scales and

native speaker performance on a communicatively oriented EAP test. *Melbourne Papers in Language Testing*, 2(1), 1-23.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll (Ed.), *Second Language Writing: Research insights for the classroom* (pp.69-87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991a). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 5-15). Westport, CT: Ablex Publishing.

Hamp-Lyons, L. (1991b). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 241-278). Westport, CT: Ablex Publishing.

Hamp-Lyons, L. (1995). Rating nonnative writing: the trouble with holistic scoring. *TESOL Quarterly*, 29(4), 759-762.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, 8(1), 5-16.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the Dynamics of Second Language Writing* (pp. 162-189). Cambridge: Cambridge University Press.

Hamp-Lyons, L. and Kroll, B. (1996). Issues in ESL writing assessment: an overview. *College ESL*, 6(1), 52-72.

Hamp-Lyons, L. and Kroll, B. (1997). *TOEFL 2000--Writing: Composition, community and assessment* (Monograph series No. RM-96-5). Princeston, NJ: Educational Testing Service.

Hamp-Lyons, L. and Prochnow, S. (1991). Combining holistic scoring with diagnostic feedback in a large-scale writing assessment. *Language Testing Update*, 9,12-15.

Harris, D. P. (1969). *Testing English as a Second Language*. New York: McGraw-Hill.

Harris, J., Laan, S. and Mossenson, L. (1988). Applying partial credit analysis to the construction of narrative writing tests. *Applied Measurement in Education*, 1(4), 335-346.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy and S. Ransdell (Eds.), *The Science of Writing* (pp. 1-27). NJ: Lawrence Erlbaum Associates.

Hayes, J. and Flower, L. (1980). Identifying the organisation of writing processes. In L. W. Gregg and E. R. Steinberg (Eds.), *Cognitive Processes in Writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hedge, T. (1998). *Writing*. Oxford: Oxford University Press.

Henning, G. (1987). *A Guide to Language Testing: development, evaluation, research*. Boston, Mass.: Heinle and Heinle Publishers.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing*, 13(1), 53-61.

Higgs, T. V. (1984). *Teaching for Proficiency, the Organising Principle*. Lincolnwood, IL: National Textbook Company.

Hill, K. (1995). Scales and tests: a case study. *Melbourne Papers in Language Testing*, 4(2), 43-59.

Hill, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds.), *Current Developments and Alternatives in Language assessment-Proceedings of LTRC 96* (pp. 275-290). Jyvaskyla: University of Jyvaskyla and University of Tampere.

Hill, K. and Storch, N. (1994). Analytic rating scales: how diagnostic are they? *Melbourne Papers in Language Testing*, 3(1), 50-65.

Ho Fong Wan Kam, B. (1985). *A Dairy Study of Teaching EFL through English and Chinese to Early Secondary School Students in Remedial English Classrooms*. Unpublished master's thesis, Chinese University of Hong Kong.

Hoetker, J. and Brossell, G. (1986). A procedure for writing content-fair essay examination topics for large-scale writing assessments. *College Composition and Communication*, 37(3), 328-335.

Horowitz, D. (1991). ESL writing assessments: contradictions and resolutions. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 71-86). Westport, CT: Ablex Publishing.

Huot, B. (1990a). Reliability, validity, and holistic scoring: what we know and what we need to know. *College Composition and Communication*, 41(2), 201-213.

Huot, B. (1990b). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research*, 60(2), 237-263.

Hwang, P. (1930). *Errors and Improvement in Rating English Compositions by means of a Composition Scale*. New York: Teachers College, Columbia University.

Hyland, K. (2002). *Teaching and Researching Writing*. London: Longman.

Hyland, K. (2003). *Second Language Writing*. Cambridge: Cambridge University Press.

Ingram, D., E. (1990). The Australian Second Language Proficiency Ratings (ASLPR). *AILA Review*, 7, 46-61.

Ingram, D. E. (1995). Scales. *Melbourne Papers in Language Testing*, 4(2), 12-29.

Ingram, D. E. and Wylie, E. (1991). Developing proficiency scales for communicative assessment. *Language and Language Education: Working Papers of the National Languages Institute of Australia*, 1(1): 31-60. [ERIC FL019252 / ED342209].

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F. and Hughey, J. B. (1981). *Testing ESL Composition: A practical approach*. Rowley, Mass: Newbury House Publishers.

Jensen, G. H. and DiTiberio, J. K. (1989). *Personality and the Teaching of Composition*. Norwood, NJ: Ablex.

Johns, A. M. (1990). L1 composition theories: implications for developing theories of L2 composition. In B. Kroll (Ed.), *Second Language Writing: Research insights for the classroom* (pp. 24-36). Cambridge: Cambridge University Press.

Johnson, C. (1985). The emergence of present perfect verbs forms: Semantic influences on selective imitation. *Journal of Child Language*, 12, 325-352.

Johnson, D. (1992). *Approaches to Research in Second Language Learning*. New York: Longman.

Kenyon, D. (1997). Further research on the efficacy of rater self-training. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds.), *Current Developments and Alternatives in Language Assessment-Proceedings of LTRC 96* (pp. 257-273). Jyvaskyla: University of Jyvaskyla and University of Tampere.

Klein, J. and Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing*, 10, 134-148.

Kleinmann, H. H. (1977). Avoidance behavior in adult second language acquisition. *Language Learning*, 27(1), 93-107.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3-31.

Kroll, B. and Reid, J. (1994). Guidelines for designing writing prompts: clarifications, caveats, and cautions. *Journal of Second Language Writing*, 3(3), 231-255.

Labov, W. (1969). *The Logic of Non-Standard English*. Urbana, IL: National Council of Teachers of English.

Lado, R. (1961). *Language Testing: The construction and use of foreign language tests*. New York: McGraw-Hill.

Lee, E.-J. (1997). *Acquisition of Tense and Aspect by Two Korean Speakers of English: A longitudinal study*. Unpublished master's thesis, University of Hawai'i, Manoa.

Leeson, R. (1975). *Fluency and Language Teaching*. London: Longman.

Lennon, P. (1990). Investigating fluency in EFL: a quantitative approach. *Language Learning*, 40, 387-417.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper and L. Odell (Eds.), *Evaluating Writing* (pp. 33-69). New York: National Council of Teachers of English.

Lumley, T. (1995). The judgements of language-trained raters and doctors in a test of English for health professionals. *Melbourne Papers in Language Testing*, 4(1), 74-98.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19(3), 246-276.

Lumley, T. and McNamamra, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12(1), 54-71.

Lynch, B. K. (2003). *Language Assessment and Programme Evaluation*. Edinburgh: Edinburgh University Press.

Martin, J. R., Christie, F. and Rothery, J. (1987). Social processes in education. *The Teaching of English: Journal of the English Teachers Association of New South Wales*, 53, 3-22.

Matthews, M. (1990). The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations. *ELT Journal*, 44(2), 117-121.

Mayer, R. E. (1984). *Thinking, Problem Solving and Cognition*. New York: W. H. Freeman and Company.

McNamara, T. F. (1990). *Assessing the Second Language Proficiency of Heatlh Professionals*. Unpublished doctoral dissertation, University of Melbourne, Melbourne.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

Meara, P. (1978). Learners' word associations in French. *Interlanguage Studies Bulletin*, 3, 192-211.

Meara, P. (1984). The study of lexis in interlanguage. In A. Davies, C. Criper and A. P. R. Howatt (Eds.), *Interlanguage* (pp. 225-239). Edinburgh: Edinburgh University Press.

Messick, S. (1988). The once and future uses of validity: assessing the meaning and consequences of measurement. In H. Wainer and H. I. Braun (Eds.), *Test Validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.,

pp. 13-103). New York: Macmillan.

Michael, W. B., Cooper, T., Shaffer, P. and Wallis, E. (1980). A comparison of the reliability and validity of ratings of student performance on essay examinations by professors of English and by professors in other disciplines. *Educational and Psychological Measurement*, 40, 183-195.

Miller, W. H. (1995). *Alternative Assessment Techniques for Reading and Writing*. New York: The Centre for Applied Research in Education.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23, 5-12.

Murphy, S. and Smith, M. A. (1991). *Writing Portfolios: A bridge from teaching to assessment*. Markham, Canada: Pippin.

Murphy, S. E. A. (1993). Survey of postsecondary writing assessment practices: Report to the CCCC Executive Committee.

Myers, I. (1962). *The Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Myers, I. (1987). *Introduction to Type* (4th ed.). Palo Alto, CA: Consulting Psychologists Press.

Nattinger, J. R. and DeCarrico, J. S. (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

North, B. (1994). Itembanker: a testing tool for language teachers. *Language Testing Update*, 16, 85-97.

North, B. (1995). Scales of language proficiency. *Melbourne Papers in Language Testing*, 4(2), 60-111.

North, B. (2000a). Defining a flexible common measurement scale: descriptors for self and teacher assessment. In G. Ekbatani and H. Pierson (Eds.), *Learner-Directed Assessment in ESL* (pp. 13-47). Mahwah, NJ: Lawrence Erlbaum Associates.

North, B. (2000b). *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang Publishing.

Nunan, D. (1999). *Second Language Teaching and Learning*. Boston, Mass: Heinle and Heinle Publishers.

Nystrand, M. (1987). The role of context in written communication. In R. Horowitz and S. J. Samuels (Eds.), *Comprehending Oral and Written Language* (pp. 197-214). San Diego, CA: Academic Press.

Odlin, T. (1989). *Language Transfer*. Cambridge: Cambridge University Press.

O'Loughlin, K. (1992). Do English and ESL teachers rate essays differently? *Melbourne Papers in Language Testing*, 1(2), 19-44.

Parkinson, B., Benson, C. and Jenkins, M. (2003). Learner diary research with 'Cambridge' examination candidates. *Edinburgh Working Papers in Applied Linguistics*, 12, 45-63.

Pollitt, A. and Hutchinson, C. (1987). Calibrating graded assessments: rasch partial credit analysis of performance in writing. *Language Testing*, 4(1), 72-92.

Pollitt, A. and Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic and N. Saville (Eds.), *Performance Testing, Cognition and Assessment* (pp. 74-89). Cambridge: Cambridge University Press.

Raimes, A. (1983). *Techniques in Teaching Writing*. Oxford: Oxford American English.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes*, 9, 109-121.

Richards, J. C. (2002). Accuracy and fluency revisited. In E. Hinkel and S. Fotos (Eds.), *New Perspectives on Grammar Teaching in Second Language Classrooms* (pp. 35-50). Mahwah, NJ: Lawrence Erlbaum Associates.

Richards, J. C., Platt, J. and Platt, H. (1992). *Longman Dictionary of Language Teaching and Applied Linguistics* (2nd ed.). London: Longman.

Richards, J. C. and Rodgers, T. S. (1986). *Approaches and Methods in Language Teaching: A description and analysis.* Cambridge: Cambridge University Press.

Ruth, L. (1982). *Properties of Writing Tasks: A study of alternative procedures for holistic writing assessment* (Bay Area Writing Project No. NIE Final Report G-809-0034). Berkeley, Calif.

Sasaki, M. and Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing*, 16(4), 457-478.

Scarino, A. (1995). Language scales and language tests: development in LOTE. *Melbourne Papers in Language Testing*, 4(2), 30-42.

Schoonen, R., Vergeer, M. and Eiting, M. (1997). The assessment of writing ability: expert readers versus lay readers. *Language Testing*, 14(2), 157-184.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, 18(3), 303-325.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1, 147-170.

Shohamy, E. (1997). Testing methods, testing consequences: are they ethical? Are they fair? *Language Testing*, 14(3), 340-349.

Silva, T. (1990). Second language composition instruction: developments, issues, and directions in ESL. In B. Kroll (Ed.), *Second Language Writing: Research*

*insights for the Classroom* (pp. 11-23). Cambridge: Cambridge University Press.

Smith, C. S. (1980). The acquisition of time talk: Relations between child and adult grammar. *Journal of Child Language*, 7, 263-278.

Song, B. and Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing*, 5(2), 163-182.

Storch, N. (1993). The development and validation of a writing test and an analytic scoring scheme used in the ESL program. *Melbourne Papers in Language Testing*, 2(2), 1-34.

Swales, J. (1990). *Genre Analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. and Horowitz, D. (1988). Genre-based approaches to ESL and ESP materials. Paper presented at the handout from a paper presented at the 22nd Annual Convention of TESOL, Chicago, IL.

Tedick, D. J. (1990). ESL writing assessment: subject-matter knowledge and its impact on performance. *English for Specific Purposes*, 9, 123-143.

Torrance, H. (1998). Learning from research in assessment: a response to writing assessment--raters' elaboration of the rating task. *Assessing Writing*, 5(1), 31-37.

Towell, R., Hawkins, R. and Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17, 84-119.

Tribble, C. (1996). *Writing*. Oxford: Oxford University Press.

Trim, J. L. M. (1978). *Some Possible Lines of Development of an Overall Structure for a European Unit/Credit Scheme for Foreign Language Learning by Adults*. Strasbourg: Council of Europe.

Turner, C. E. and Upshur, J. A. (2002). Rating scales derived from student samples: effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, 36(1), 49-70.

Underhill, N. (1987). *Testing Spoken Language: A handbook of oral testing techniques*. Cambridge: Cambridge University Press.

Upshur, J. A. and Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, 49(1), 3-12.

Vaughan, C. (1991). Holistic assessment: what goes on in the raters' minds? In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 111-125). Westport, CT: Ablex Publishing.

von Stutterheim, C. W. and Klein, W. (1987). A concept-oriented approach to second language studies. In C.W. Pfaff (Ed). *First and Second Language Acquisition Processes* (pp. 191-205). Cambridge, MA: Newbury House.

Warschauer, M. and Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research*, 10(2), 157-180.

Webb, E. (1915). *Character and Intelligence: An attempt of an exact study of character*. Cambridge: Cambridge University Press.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C. J. (1990). *Communicative Language Testing*. New York: Prentice Hall.

Weir, C. J. (1993). *Understanding and Developing Language Tests*. Hemel Hempstead, England: Phoenix ELT.

White, E. M. (1994). *Teaching and Assessing Writing: Recent advances in understanding evaluating, and improving student performance* (2nd ed.). San Francisco, Calif: Jossey-Bass Publishers.

White, E. M. (1995). An apologia for the timed impromptu essay test. *College Composition and Communication*, 46(1), 30-45.

White, J. (1989). Children's writing: some findings from data collected longitudinally. *Research Papers in Education*, 4(2), 53-78.

Wolfe-Quintero, K., Inagaki, S. and Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of fluency, accuracy and complexity*. Hawaii: Second Language Teaching and Curriculum Centre, University of Hawaii at Manoa.

**From website**

the IELTS http://www.ielts.org

the ISLPR http://www.gu.edu.au/centre/call

Lotus ScreenCam http://www.lotus.com/products/screencam.nsf

Scales for the Test of Written Examination (TWE) of the ETS http://www.toefl.org

Test of Writing Proficiency (TWP) http://www.teps.or.kr

# APPENDICES

**Appendix 1. The Model of writing ability** (Grabe & Kaplan, 1996: 220-1)

I. Linguistic knowledge

A. Knowledge of the written code
1. Orthography
2. Spelling
3. Punctuation
4. Formating conventions (margins, paragraphing, spacing, etc.)

B. Knowledge of phonology and morphology
1. Sound/letter correspondences
2. Syllables
(1) onset
(2) rhyme/rhythm
(3) coda
3. Morpheme structure (word-part knowledge)

C. Vocabulary
1. Interpersonal words and phrases
2. Academic and pedagogical words and phrases
3. Formal and technical words and phrases
4. Topic-specific words and phrases
5. Non-literal and metaphoric language

D. Syntactic/structural knowledge
1. Basic syntactic patterns
2. Preferred formal writing structures (appropriate style)
3. Tropes and figures of expression
4. Metaphors/similes

E. Awareness of differences across languages

F.Awareness of relative proficiency in different languages and registers

II.Discourse knowledge

A. Knowledge of intrasentential and intersentential marking devices (cohesion, syntactic parallelism)

B. Knowledge on informational structuring (topic/comment, given/new, theme/rheme, adjacency pairs)

C. Knowledge of semantic relations across clauses

D. Knowledge to recognise main topics

E. Knowledge of genre structure and genre constraints

F. Knowledge of organising schemes (top-level discourse structure)

G. Knowledge of inferencing (bridging, elaborating)

H. Awareness of differences in features of discourse structuring across languages and cultures

I. Awareness of different proficiency levels of discourse skills in different languages

II. Sociolinguistic knowledge

A. Functional uses of written language
1. Apologise
2. Deny
3. Complain
4. Threaten
5. Invite
6. Agree
7. Congratulate

8. Request
9. Direct
10. Compliment

B. Application and interpretable violation of Gricean maxims
C. Register and situational parameters
1. Age of writer
2. Language used by writer (L1, L2…)
3. Proficiency in language used
4. Audience considerations
5. Relative status of interactants (power/politeness)
6. Degree of formality (deference/solidarity)
7. Degree of distance (detachment/involvement)
8. Topic of interaction
9. Means of writing (pen/pencil, computer, dictation, shorthand)
10. Mean of transmission (single page/book/read aloud/printed)

D. Awareness of sociolinguistic differences across languages and cultures
E. Self-awareness of roles of register and situational parameters

**Appendix 2. Questionnaire I**

## A Questionnaire on English teachers' Use of a Rating Scale for Writing Assessment for Korean High School Students

I am pleased to get in touch with you again. My name is Jyi-yeon Yi and I'm on a PhD research programme in the field of applied linguistics at the University of Edinburgh in the U.K. at present. As you know, I am writing to ask you to complete the questionnaire below.

While I appreciate that you are very busy with your work, I would be very grateful if you would complete this questionnaire.

**Before you start**, please **read the tips for responding below** and try to follow them.

Should you have any queries about this survey, please do not hesitate to contact me by e-mail (jyiyeon@ling.ed.ac.uk.).

Thank you in advance for responding to the survey

Yours sincerely,

Jyi-yeon Yi

### Tips for responding

- Please note that this questionnaire is on writing assessment for Korean high school students, and that by the term "writing assessment" referred to in the questionnaire, I mean free writing, where candidates on the writing test are expected to write freely for a given task, which is usually between half a page to one or more pages long.
- Please remember that a rating scale in the questionnaire is not intended to be an abstract concept. Rather, it is supposed to be a substantial yardstick on which assessment categories (which measure aspects such as grammar, content, style, or organization, etc.) are differentiated and described, ascending or descending according to the assessment grades, such as 1-2-3-4 or beginner-intermediate-advanced, etc.
- You are asked to **answer all the questions from Question 1 to Question 8.** After Question 9 you will be invited to go to a specific question according to the choice you made in the previous question. Please make sure that you follow the instructions provided **next to each choice** to indicate which question you should answer next.
- Please make sure that you only choose one of the options in each question.

Q1. Sex

① Male

② Female

Q2. Age

① Under 30

② 31-40

③ 41- 50

④ 51-60

⑤ Over 61

Q3. Academic career

① Completion of undergraduate

② Currently studying for postgraduate for Master degree

③ Completion of postgraduate (Master degree)

④ Currently studying for postgraduate for Doctorate degree

⑤ Completion of postgraduate (PhD degree)

Q4. Career

① A Korean regular teacher at a foreign language high school

② A Korean part-time or definite-period teacher at a foreign language high school

③ An English- native-speaker regular teacher at a foreign language high school

④ An English-native speaker part-time lecturer at a foreign language high school

Q5. Service area

① Seoul

② Kyungki

③ Kangwon / Choongchung

④ Cheonbuk / Cheonnam

⑤ Kyungbuk / Kyungnam

Q6. Length of time for which you have held the current career position

① Under 1 year

② Between 1 and 2 years

③ Between 2 and 5 years

④ Between 5 and 10 years

⑤ Over 10 years

Q7. Subject that you teach at present

① English I/II
② Reading
③ Writing
④ Conversation
⑤ Other

Q8. How often do you invite your students to write (free-write) and accordingly assess them?

① More than twice a week (⇨Please go to Q11 if you chose ①)
② About once a week (⇨ to Q11)
③ About once a month (⇨ to Q11)
④ About once or twice a term (⇨ to Q11)
⑤ Never (⇨ to Q9)

Q9. If you do not have writing session in your class, why is it?

① Because there is a specific course where writing is mainly handled (⇨ to Q10)
② Because students' ability to write freely is not developed enough to be assessed (⇨ to Q10)
③ Because it takes a great deal of time to assess students' writing (⇨ to Q10)
④ Because there is no reliable assessment standard (⇨ to Q 10)
⑤ Other (⇨ to Q10)

Q10. Is there any particular rating scale for writing assessment recommended to you and other English teachers through the ministry of education, school or in-service training for teachers?

① Yes (⇨ to Q23)
② No (⇨ to Q23)

Q11. Is there any particular rating scale for writing assessment recommended to you and other English teachers through the ministry of education, school or in-service training for teachers?

① Yes (⇨ to Q12)
② No (⇨ to Q14)

Q12. Do you use the recommended rating scale in assessing?

① Every time I do writing assessment (⇨ to Q23)
② Often (⇨ to Q23)
③ Occasionally (⇨ to Q13)
④ Never (⇨ to Q13)

Q13. Why do you not use the recommended rating scale?

① Because the content of the rating scale is not suitable for your students' English writing ability and the curriculum for high school (⇨ to Q14)
② Because the assessing categories and grades of the rating scale are so complex that it is inconvenient to use it (⇨ to Q14)
③ Because the terms in the rating scale are so vague to interpret that it is difficult to apply it in practice (⇨ to Q14)
④ Because there is another rating scale that I have got accustomed to using (⇨ to Q14)
⑤ Other (⇨ to Q14)

Q14. How do you assess your students' writing?

① According to my judgement (or insight), rather than using a formal rating scale (⇨ to Q15)
② Using a rating scale that I have developed for myself or with my colleagues (⇨ to Q18)
③ Using other existing rating scale (⇨ to Q21)
④ Other (⇨ to Q23)

Q15. How do you assess your students' writing when you assess according to your judgement rather than a formal rating scale?

① According to whether a student in question has accomplished a writing task given by me (⇨ to Q16)
② According to the overall impression that I have got from reading it (⇨ to Q16)
③ Mainly according to whether handwriting is legible and tidy or whether its length (quantity) is enough (⇨ to Q16)
④ Mainly according to the extent to which it is organized (⇨ to Q16)
⑤ Other (⇨ to Q16)

Q16. Why do you not use any rating scale in assessing?

① Because it is difficult to obtain rating scales (⇨ to Q17)

② Because there is no suitable rating scale among existing rating scales (⇨ to Q17)
③ Because I find using rating scales too complicated (⇨ to Q17)
④ Because I find it enough to assess writing using my own insight (⇨ to Q17)
⑤ Other (⇨ to Q17)

Q17. What do you think the most significant problem is, if any, when you do not use any formal rating scales in writing assessment?

① I am worried about whether it can achieve consistency of obtained results. (⇨ to Q23)
② I am worried about whether it assesses aspects that should be measured in writing ability. (⇨ to Q23)
③ I do not have any particular problems. (⇨ to Q23)
④ Other (⇨ to Q23)

Q18. Why do you use your own rating scale?

① Because it is difficult to obtain existing rating scales (⇨ to Q19)
② Because I find existing rating scales too complex to use (⇨ to Q19)
③ Because there is no suitable one among existing rating scales for my students' English writing ability (⇨ to Q19)
④ Because I find it convenient and effective to use a rating scale which I have developed (⇨ to Q19)
⑤ Other (⇨ to Q19)

Q19. How do you proceed when developing a rating scale for yourself or with your colleagues?

① By adapting some features of one or more existing rating scales (⇨ to Q20)
② By reflecting aspects which you find important in writing (⇨ to Q20)
③ By basing it on students' previous writing samples (⇨ to Q20)
④ By reflecting lesson goals of the course (⇨ to Q20)
⑤ Other (⇨ to Q20)

Q20. What do you think the most significant problem is, if any, with the rating scale that you have developed?

① I feel it is too time-consuming and requires too much effort to develop. (⇨ to Q23)
② I am worried about whether it can achieve consistency of obtained results. (⇨ to Q23)
③ I am worried about whether it assesses aspects that should be measured in writing

ability. (⇨ to Q23)

④ I do not have any particular problems. (⇨ to Q23)

⑤ Other (⇨ to Q23)

Q21. How do you obtain the existing rating scale that you use?

① By referring to relevant books (⇨ to Q22)

② By browsing the Internet (⇨ to Q22)

③ Through in-service training for teachers (⇨ to Q22)

④ Through a teacher's society (⇨ to Q22)

⑤ Other (⇨ to Q22)

Q22. What do you think the most significant problem is, if any, with the existing rating scale that you use?

① It takes long time to get accustomed to using the rating scale as its content is complex and is not clear enough to understand. (⇨ to Q23)

② The content of the rating scale is not completely suitable to my students' English writing ability. (⇨ to Q23)

③ There are some aspects in it that are not suitable to the curriculum. (⇨ to Q23)

④ I do not have any particular problems. (⇨ to Q23)

⑤ Other (⇨ to Q23)

Q23. Do you think it is desirable to develop rating scales for classroom writing assessment, specifically for Korean high school students?

① Definitely disagree (⇨ to Q24)

② Slightly disagree (⇨ to Q24)

③ Slightly agree (⇨ to Q25)

④ Definitely agree (⇨ to Q25)

Q24. Why do you not think it is desirable to develop a rating scale for classroom writing assessment, specifically for Korean high school students?

① Because little emphasis is still put on writing skill

② Because less emphasis is put on writing assessment than on reading or listening assessment

③ Because I find it enough to use existing rating scales

④ Because I find it enough to use my own rating scale or judgement

⑤ Other

Q25. Why do you think it is desirable to develop a rating scale for classroom writing assessment, specifically for Korean high school students?

① Because I find existing rating scales too complex to use

② Because the content of existing rating scales is not suitable for Korean high school students' English writing ability or the curriculum

③ Because the terms or distinctions between assessing grades in existing rating scales are not clear enough so that it is difficult to apply it in practice.

④ Because it should be developed some time in the future

⑤ Other

Thank you very much.

**Appendix 3. The results of Questionnaire I survey**

## A Questionnaire on English teachers' Use of a Rating Scale for Writing Assessment for Korean High School Students

**Jyi-yeon Yi** (jyiyeon@ling.ed.ac.uk) University of Edinburgh, U.K.

Period : 21 May 2003 ~ 31 May 2003

Online URL : http://www.research.joongang.com/survey.php?id=03-fine-615

Total Number of Respondents : **35**

1. Sex

| | | |
|---|---|---|
| Male | 26 | 74% |
| Female | 9 | 26% |
| No Answer | 0 | 0% |

2. Age

| | | |
|---|---|---|
| Under 30 | 6 | 17% |
| 31-40 | 16 | 46% |
| 41- 50 | 13 | 37% |
| 51-60 | 0 | 0% |
| Over 61 | 0 | 0% |
| No Answer | 0 | 0% |

3. Academic career

| | | |
|---|---|---|
| Completion of undergraduate | 15 | 43% |
| Currently studying for postgraduate for Master degree | 5 | 14% |
| Completion of postgraduate (Master degree) | 9 | 26% |
| Currently studying for postgraduate for Doctorate degree | 5 | 14% |
| Completion of postgraduate (PhD degree) | 1 | 3% |
| No Answer | 0 | 0% |

4. Career

| | | |
|---|---|---|
| A Korean regular teacher at a foreign language high school | 35 | 100% |
| A Korean part-time or a definite-period teacher at a foreign language high school | 0 | 0% |

| | | |
|---|---|---|
| An English- native-speaker regular teacher at a foreign language high school | 0 | 0% |
| An English-native speaker part-time lecturer at a foreign language high school | 0 | 0% |
| No Answer | 0 | 0% |

5. Service area

| | | |
|---|---|---|
| Seoul | 8 | 23% |
| Kyungki | 12 | 34% |
| Kangwon / Choongchung | 1 | 3% |
| Cheonbuk / Cheonnam | 0 | 0% |
| Kyungbuk / Kyungnam | 14 | 40% |
| No Answer | 0 | 0% |

6. Length of time for which you have held the current career position

| | | |
|---|---|---|
| Under 1 year | 3 | 9% |
| Between 1 and 2 years | 2 | 6% |
| Between 2 and 5 years | 3 | 9% |
| Between 5 and 10 years | 10 | 29% |
| Over 10 years | 17 | 49% |
| No Answer | 0 | 0% |

7. Subject that you teach at present

| | | |
|---|---|---|
| English I/II | 8 | 23% |
| Reading | 17 | 49% |
| Writing | 0 | 0% |
| Conversation | 1 | 3% |
| Other | 9 | 26% |
| No Answer | 0 | 0% |

8. How often do you invite your students to write (free-write) and accordingly assess them?

| | | |
|---|---|---|
| More than twice a week | 2 | 6% |
| About once a week | 8 | 23% |
| About once a month | 4 | 11% |
| About once or twice a term | 9 | 26% |
| Never | 12 | 34% |
| No Answer | 0 | 0% |

9. If you do not have writing session in your class, why is it?

| | | |
|---|---|---|
| Because there is a specific course where writing is mainly handled | 5 | 14% |
| Because students' ability to write freely is not developed enough to be assessed | 2 | 6% |
| Because it takes a great deal of time to assess students'writing | 2 | 6% |
| Because there is no reliable assessment standard | 2 | 6% |
| Other | 1 | 3% |
| No Answer | 23 | 66% |

10. Is there any particular rating scale for writing assessment recommended to you and other English teachers through the ministry of education, school or in-service training for teachers?

| | | |
|---|---|---|
| Yes | 2 | 6% |
| No | 10 | 29% |
| No Answer | 23 | 66% |

11. Is there any particular rating scale for writing assessment recommended to you and other English teachers through the ministry of education, school or in-service training for teachers?

| | | |
|---|---|---|
| Yes | 1 | 3% |
| No | 22 | 63% |
| No Answer | 12 | 34% |

12. Do you use the recommended rating scale in assessing?

| | | |
|---|---|---|
| Every time I do writing assessment | 0 | 0% |
| Often | 0 | 0% |
| Occasionally | 1 | 3% |
| Never | 1 | 3% |
| No Answer | 33 | 94% |

13. Why do you not use the recommended rating scale?

| | | |
|---|---|---|
| Because the content of the rating scale is not suitable for your students' English writing ability and the curriculum for high school | 0 | 0% |
| Because the assessing categories and grades of the rating scale are so complex that it is inconvenient to use it | 0 | 0% |

| | | |
|---|---|---|
| Because the terms in the rating scale are so vague to interpret that it is difficult to apply it in practice | 0 | 0% |
| Because there is another rating scale that I have got accustomed to using | 1 | 3% |
| Other | 1 | 3% |
| No Answer | 33 | 94% |

14. How do you assess your students' writing?

| | | |
|---|---|---|
| According to my judgement (or insight), rather than using a formal rating scale | 11 | 31% |
| Using a rating scale that I have developed for myself or with my colleagues | 8 | 23% |
| Using other existing rating scale | 3 | 9% |
| Other | 1 | 3% |
| No Answer | 12 | 34% |

15. How do you assess your students' writing when you assess according to your judgement rather than a formal rating scale?

| | | |
|---|---|---|
| According to whether a student in question has accomplished a writing task given by me | 2 | 6% |
| According to the overall impression that I have got from reading it | 1 | 3% |
| Mainly according to whether handwriting is legible and tidy or whether its length (quantity) is enough | 2 | 6% |
| Mainly according to the extent to which it is organised | 6 | 17% |
| Other | 0 | 0% |
| No Answer | 24 | 69% |

16. Why do you not use any rating scale in assessing?

| | | |
|---|---|---|
| Because it is difficult to obtain a rating scale | 2 | 6% |
| Because there is no suitable rating scale among existing rating scales | 5 | 14% |
| Because I find using rating scales too complicated | 2 | 6% |
| Because I find it enough to assess writing using | 2 | 6% |

| | | |
|---|---|---|
| my own insight | | |
| Other | 0 | 0% |
| No Answer | 24 | 69% |

17. What do you think the most significant problem is, if any, when you do not use any formal rating scale in writing assessment?

| | | |
|---|---|---|
| I am worried about whether it can achieve consistency of obtained results. | 8 | 23% |
| I am worried about whether it assesses aspects that should be measured in writing ability. | 3 | 9% |
| I do not have any particular problems. | 0 | 0% |
| Other | 0 | 0% |
| No Answer | 24 | 69% |

18. Why do you use your own rating scale?

| | | |
|---|---|---|
| Because it is difficult to obtain existing rating scales | 2 | 6% |
| Because I find existing rating scales too complex to use | 0 | 0% |
| Because there is no suitable one among existing rating scales for my students' English writing ability | 2 | 6% |
| Because I find it convenient and effective to use a rating scale which I have developed | 4 | 11% |
| Other | 0 | 0% |
| No Answer | 27 | 77% |

19. How do you proceed when developing a rating scale for yourself or with your colleagues?

| | | |
|---|---|---|
| By adapting some features of one or more existing rating scales | 3 | 9% |
| By reflecting aspects which you find important in writing | 4 | 11% |
| By basing it on students' previous writing samples | 0 | 0% |
| By reflecting lesson goals of the course | 1 | 3% |
| Other | 0 | 0% |
| No Answer | 27 | 77% |

20. What do you think the most significant problem is, if any, with the rating scale that you have developed?

| | | |
|---|---|---|
| I feel it is too time-consuming and requires too much effort to develop. | 0 | 0% |
| I am worried about whether it can achieve consistency of obtained results. | 2 | 6% |
| I am worried about whether it assesses aspects that should be measured in writing ability. | 3 | 9% |
| I do not have any particular problems. | 3 | 9% |
| Other | 0 | 0% |
| No Answer | 27 | 77% |

21. How do you obtain the existing rating scale that you use?

| | | |
|---|---|---|
| By referring to relevant books | 3 | 9% |
| By browsing the Internet | 0 | 0% |
| Through in-service training for teachers | 0 | 0% |
| Through a teacher's society | 0 | 0% |
| Other | 0 | 0% |
| No Answer | 32 | 91% |

22. What do you think the most significant problem is, if any, with the existing rating scale that you use?

| | | |
|---|---|---|
| It takes long time to get accustomed to using the rating scale as its content is complex and is not clear enough to understand. | 0 | 0% |
| The content of the rating scale is not completely suitable to my students' English writing ability. | 1 | 3% |
| There are some aspects in it that are not suitable to the curriculum. | 1 | 3% |
| I do not have any particular problems. | 1 | 3% |
| Other | 0 | 0% |
| No Answer | 32 | 91% |

23. Do you think it is desirable to develop a rating scale for classroom writing assessment, specifically for Korean high school students?

| | | |
|---|---|---|
| Definitely disagree | 1 | 3% |

| | | |
|---|---|---|
| Slightly disagree | 4 | 11% |
| Slightly agree | 18 | 51% |
| Definitely agree | 12 | 34% |
| No Answer | 0 | 0% |

24. Why do you not think it is desirable to develop a rating scale for classroom writing assessment, specifically for Korean high school students?

| | | |
|---|---|---|
| Because little emphasis is still put on writing skill | 2 | 6% |
| Because less emphasis is put on writing assessment than on reading or listening assessment | 0 | 0% |
| Because I find it enough to use existing rating scales | 1 | 3% |
| Because I find it enough to use my own rating scale or judgement | 1 | 3% |
| Other | 1 | 3% |
| No Answer | 30 | 86% |

25. Why do you think it is desirable to develop a rating scale for classroom writing assessment, specifically for Korean high school students?

| | | |
|---|---|---|
| Because I find existing rating scales too complex to use | 0 | 0% |
| Because the content of existing rating scales is not suitable for Korean high school students' English writing ability or the curriculum | 6 | 17% |
| Because the terms or distinctions between assessing grades in existing rating scales are not clear enough so that it is difficult to apply it in practice. | 9 | 26% |
| Because it should be developed some time in the future | 14 | 40% |
| Other | 1 | 3% |
| No Answer | 5 | 14% |

NB. The respondents of 'No answer' were the people who did not have to answer for the question, according to the instruction.

**Appendix 4. Questionnaire II**

# A Questionnaire on Writing Tasks

Writing assessments were carried out over two sessions. One of them was to 'write a letter to your foreign friend introducing your life at school' and the other was to 'write an essay to a teacher explaining the advantages and disadvantages of living in a big city'.
I would like to ask your opinion on the two writing assessments. It would be appreciated if you would answer the questions below, following The Tips for Responding.

**Tips for Responding**

- Please note that you should only choose one of the options given in each question, unless there is a specific instruction for multiple answers to a question.
- Please note that you do not have to answer all of the questions. Please make sure that you follow the instructions provided next to each choice indicating which question you will be invited to answer next.

I. Respondents' Information:

1. Year _____
2. Class _____
3. Sex _____

II. The following questions are about the task 'write a letter to your foreign friend introducing your life at school'.

1. How difficult did you find the task?

① Fairly difficult (⇒Please go to Q2)
② Slightly difficult (⇒Q2)
③ Reasonably easy (⇒Q3)
④ Very easy (⇒Q3)

2. Please answer the following questions regarding the extent to which the following aspects influenced your performance.

2-1 Did you have any problems with the format of a letter in English?

① No problems (⇒Q2-2)
② Minor problems (⇒Q2-2)
③ Moderate problems (⇒Q2-2)

④ Serious problems (⇒Q2-2)

2-2 Did you have any problems deciding how to write about the topic?

① No problems (⇒Q2-3)

② Minor problems (⇒Q2-3)

③ Moderate problems (⇒Q2-3)

④ Serious problems (⇒Q2-3)

2-3 Did you have any problems finding an appropriate style to write a letter to a friend in English?

① No problems (⇒Q2-4)

② Minor problems (⇒Q2-4)

③ Moderate problems (⇒Q2-4)

④ Serious problems (⇒Q2-4)

2-4 Did you have any problems fully understanding the prompt?

① No problems (⇒Q3)

② Minor problems (⇒Q3)

③ Moderate problems (⇒Q3)

④ Serious problems (⇒Q3)

3. To what extent do you think the task gave a chance to show your English ability when doing the task?

① Not at all (⇒Q4)

② Slightly (⇒Q4)

③ Moderately (⇒Q4)

④ Enough (⇒Q4)

4. The audience for the task was specified: a foreign friend. How did you find this specification when you were doing the tasks?

① I found it very troublesome and it made the tasks very difficult (⇒Q5).

② I found it moderately troublesome and it made the tasks moderately difficult (⇒Q5).

③ I found it moderately useful in doing the tasks (⇒Q6).

④ I found it very useful in doing the tasks (⇒Q6).

5. Why do you think you found the audience specification difficult while doing the task?

① I found it annoying to write for the specified audience (⇒Q6).

② I found it unfamiliar to be supplied with a specific audience (⇒Q6).

③ Other (Please describe:_____) (⇒Q6)

6. How useful was it to have guidelines for the content of your writing while doing the task?

① Not useful at all (⇒Q7)

② Rarely useful (⇒Q7)

③ Moderately useful (⇒Q8)

④ Very useful (⇒Q8)

7. Why did you find the cues either never or rarely useful?

① I found myself sticking to the guidelines for the content and as a result it made the process of writing more difficult (⇒Q8).

② I had a number of things to write about and consequently I didn't want to take advantage of the cues (⇒Q8).

③ I found the guidelines inappropriate to the given topics (⇒Q8).

④ Other (Please describe:_____) (⇒Q8)

III. The following questions are about the task 'write an essay to a teacher explaining advantages and disadvantages of living in a big city'.

8. How difficult did you find the task?

① Fairly difficult (⇒Q9)

② Slightly difficult (⇒Q9)

③ Reasonably easy (⇒Q10)

④ Very easy (⇒Q10)

9. Please answer the following questions regarding the extent to which the following aspects influenced your performance.

9-1 Did you have any problems with the format of a letter in English?

① No problems (⇒Q9-2)

② Minor problems (⇒Q9-2)

③ Moderate problems (⇒Q9-2)

④ Serious problems (⇒Q9-2)

9-2 Did you have any problems deciding how to write about the topic?

① No problems (⇒Q9-3)

② Minor problems (⇒Q9-3)

③ Moderate problems (⇒Q9-3)

④ Serious problems (⇒Q9-3)

9-3 Did you have any problems finding an appropriate style for an formal essay?

① No problems (⇒Q9-4)

② Minor problems (⇒Q9-4)

③ Moderate problems (⇒Q9-4)

④ Serious problems (⇒Q9-4)

9-4 Did you have any problems fully understanding the prompt?

① No problems (⇒Q10)

② Minor problems (⇒Q10)

③ Moderate problems (⇒Q10)

④ Serious problems (⇒Q10)

10. To what extent do you think the task gave a chance to show your English ability when doing the task?

① Not at all (⇒Q11)

② Slightly (⇒Q11

③ Moderately (⇒Q11)

④ Enough (⇒Q11)

11. The audience for each task was specified: a teacher. How did you find this specification when you were doing the tasks?

① I found it very troublesome and it made the tasks very difficult (⇒Q12).

② I found it moderately troublesome and it made the tasks moderately difficult (⇒Q12).

③ I found it moderately useful in doing the tasks (⇒Q13).

④ I found it very useful in doing the tasks (⇒Q13).

12. Why do you think you found the audience specification difficult while doing the tasks?

① I found it annoying to write for the specified audience (⇒Q13).

② I found it unfamiliar to be supplied with a specific audience (⇒Q13).

③ Other (Please describe:_____) (⇒Q13)

13. How useful was it to have guidelines for the content of your writing?

① Not useful at all (⇒Q14)

② Rarely useful (⇒Q14)

③ Moderately useful (⇒Q15)

④ Very useful (⇒Q15)

14. Why did you find the cues either never or rarely useful?

① I found myself sticking to the guidelines for the content and as a result it made the process of writing more difficult (⇒Q15).

② I had a number of things to write about and consequently I didn't want to take advantage of the cues (⇒Q15).

③ I found the guidelines inappropriate to the given topics (⇒Q15).

④ Other (Please describe:_____) (⇒Q15)

IV. The following questions are about both tasks.

15. How useful did you find the specified minimum length (200words)?

① It was useful because it helped me to know how much I should write (⇒Q16).

② It did not matter since I usually write as much as the required quantity (⇒Q16).

③ I found myself paying attention to quantity rather than quality of writing (⇒Q16).

④ I found it very unusual and difficult to write as much as the required quantity (⇒Q16).

⑤ Other (Please describe:_____) (⇒Q16).

16. How did you feel about having only one prompt for each writing task?

① I didn't show my English ability because the given prompt for each writing task was difficult and there were no alternatives (⇒Q17).

② Although irritated that there were no alternatives, this barely affected my writing (⇒Q17).

③ I liked this as all the candidates were asked to write on an identical prompt and consequently this would help them to be assessed fairly (⇒Q17).

④ Other (Please specify:_____) (⇒Q17)

17. What kind of feedback do you want for your writing?

① A single score (e.g., A, B+, 5 points, 70 points)

② A total score plus individual scores for each writing feature (e.g., Grammar-5, Organisation-5, Content-4)

③ Scores plus a teacher's comment on your writing

④ Other (Please describe:_____)

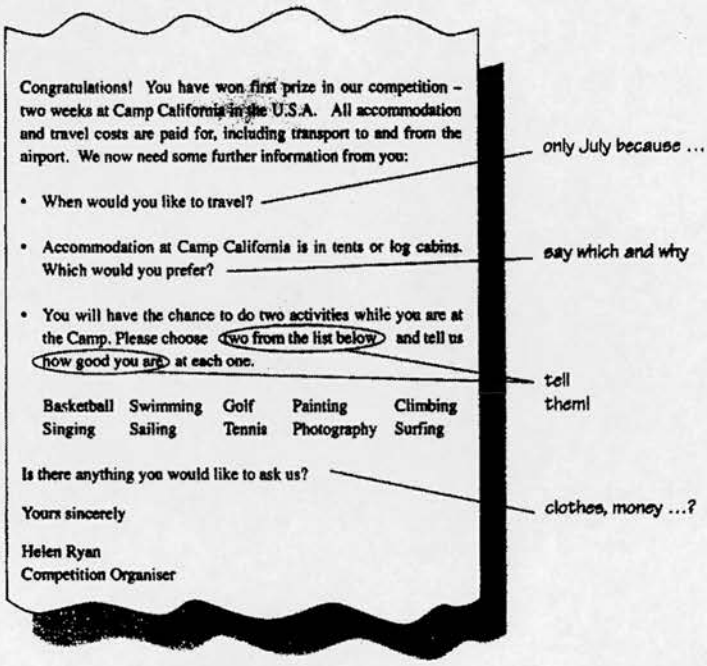
★Thank you very much★

**Appendix 5. An example of the FCE writing test** (FCE Handbook, 2001: 21)

## Part 1

You must answer this question.

1. You recently entered a competition and have just received this letter from the organiser. Read the letter, on which you have made some notes. Then, using all the information in your notes, write a suitable reply.




Write a letter of between 120 and 180 words in an appropriate style on the opposite page. Do not write any postal addresses.

## Part 2

Write an answer to one of the questions 2-5 in this part. Write your answer in 120-180 words in an appropriate style on the opposite page. Put the questions number in the box.

2. Your English class is going to make a short video about daily life at your school. Your teacher has asked you to write a report, suggesting which lesson and other activities should be filmed, and why.

Write your report.

3. Your have recently had a class discussion about shopping. Now your English teacher has asked you to write a composition, giving your opinions on the following statement:

*Shopping is not always enjoyable.*

Write your composition.

4. Last month, you enjoyed helping at a pop concert and your pen friend, Kim, wants to hear about your experience. Write a letter to Kim, describing what you did to help and explaining what you particularly liked about the experience.
Write your letter. Do not write any postal addresses.

5. Answer one of the following two questions based on your reading of one of these set books. Write (a) or (b) as well as the number 5 in the question box, and the title of the book next to the box.

*Best Detective Stories of Agatha Christie* – Longman Fiction
*The Old Man and the Sea* – Ernest Hemingway
*Cry Freedom* – John Briley
*Wuthering Heights* – Emily Bronte
*A Window on the Universe* – Oxford Bookworm Collection

Either (a) 'Sometimes the bad characters in a story are more interesting than the good ones'. Is this true of the book you have read? Write a composition, explaining your views with reference to the book or one of the short stories you have read.

Or (b) 'This is such a marvellous book you will want to read it again'. Write an article for your college magazine, saying whether you think this statement is true of the book or one of the short stories you have read.

**Appendix 6. The FCE general rating scheme for writing assessment** (FCE handbook, 2001: 19)

| | |
|---|---|
| Band 5 | Full realisation of the task set.<br>● All content points included with appropriate expansion.<br>● Wide range of structure and vocabulary within the task set.<br>● Minimal errors, perhaps due to ambition; well-developed control of language.<br>● Ideas effectively organised, with a variety of linking devices.<br>● Register and format consistent appropriate to purpose and audience.<br>Fully achieves the desired effects on the target reader |
| Band 4 | Good realisation of the task set.<br>● All major content points included; possibly one or two minor omissions.<br>● Good range of structure and vocabulary within the task set.<br>● Ideas clearly organised, with suitable linking devices.<br>● Register and format on the whole appropriate to purpose and audience.<br>Achieves the desired effect on the target reader. |
| Band 3 | Reasonable achievement of the task set.<br>● All major content points included; some minor omissions.<br>● Adequate range of structure and vocabulary, which fulfils the requirements of the task.<br>● Ideas adequately organised, with simple linking devices.<br>● Reasonable, if not always successful attempt at register and format appropriate to purpose and audience.<br>Achieves, on the whole, the desired effect on the target reader. |
| Band 2 | Task set attempted but not adequately achieved.<br>● Some major content points inadequately covered or omitted, and/or some irrelevant material.<br>● Limited range of structure and vocabulary.<br>● A number of errors, which distract the reader and may obscure communication at times.<br>● Ideas inadequately organised; linking devices rarely used.<br>● Unsuccessful/inconsistent attempts at appropriate register and format.<br>Message not clearly communicated to the target reader. |
| Band 1 | Poor attempt at the task set.<br>● Notable content omissions and/or considerable irrelevance, possibly due to misinterpretation of task set.<br>● Narrow range of vocabulary and structure.<br>● Frequent errors which obscure communication; little evidence of language control.<br>● Lack of organisation, or linking devices.<br>● Little or no awareness of appropriate register and format.<br>Very negative effect on the target reader. |
| Band 0 | Achieves nothing: too little language for assessment (fewer than 50 words) or totally irrelevant or totally illegible. |

[Originally No Table No.]

## Appendix 7. The modified FCE general rating scheme

| | Content | Accuracy | Range | Organisation and Cohesion | Appropriacy of Register and Format | Target Reader |
|---|---|---|---|---|---|---|
| Band 6 | **All content** points included with appropriate expansion | Minimal errors, perhaps due to ambition; **well-developed control** of language | **Wide range** of structure and vocabulary within the task set | Ideas **effectively organised**, with a **variety of linking devices** | Register and format **consistently appropriate** to purpose and audience | **Fully achieves** the desired effect on the target reader |
| Band 5 | **All major** content points included; possibly **one or two minor omissions** | **Generally accurate,** errors occur mainly when attempting more complex language | **Good range** of structure and vocabulary within the task set | Ideas **clearly organised,** with **suitable linking** devices | Register and format **on the whole appropriate** to purpose and audience | **Achieves** the desired effect on the target reader |
| Band 4 | All major content points included; **some minor omissions** | **A number of errors** may be present, but they do **not impede communication** | **Adequate range** of structure and vocabulary, which fulfils the requirements of the task | Ideas **adequately organised,** with **simple linking** devices | **Reasonable**, if not always successful attempt at register and format appropriate to purpose and audience | **Achieves, on the whole,** the desired effect on the target reader |
| Band 3 | Some major content points **inadequately covered** or omitted, and/or some irrelevant material | A number of errors, which **distract the reader** and any obscure communication at times | **Limited range** of structure and vocabulary | Ideas **inadequately organised**; linking devices **rarely used** | **Unsuccessful/ inconsistent attempts** at appropriate register and format | Message **not clearly communicated** to the target reader |
| Band 2 | **Notable content omissions** and/or considerable irrelevance, possibly due to misinterpretation of task set | Frequent errors which obscure communication; **little evidence of language control** | **Narrow range** of vocabulary and structure | **Lack** of organisation, or linking devices | **Little or no awareness** of appropriate register and format | **Very negative effect** on the target reader |
| Band 1 | Achieves nothing: too little language for assessment (fewer than 50 words) or totally irrelevant or totally illegible | | | | | |

## Appendix 8. Coding sheet

| | | | | | |
|---|---|---|---|---|---|
| (1) Accuracy | (1.1)Intelligible | (1.1.1) Verb | (1.1.1.1)Distinction between finite and non-finite verbs/omission or repetition of finite verbs (C) | | |
| | | | (1.1.1.2)Distinction between verb types (C) | | |
| | | (1.1.2)Tense (C) | | | |
| | | (1.1.3) Indication of quantity in nouns and Articles (C) | | | |
| | | (1.1.4)Agreement(C) | | | |
| | | (1.1.5)Conjunctions and Relatives | (1.1.5.1)Syntactic error (C) | | |
| | | | (1.1.5.2)Semantic error (C) | | |
| | | (1.1.6)Distinctions between word classes | (1.1.6.1)*because there for example* (C) | | |
| | | | (1.1.6.2)Other(C) | | |
| | | (1.1.7)Voice and Participles (C) | | | |
| | | (1.1.8)Prepositions and Particles | (1.1.8.1)Syntactic error (C) | | |
| | | | (1.1.8.2)Semantic error (C) | | |
| | | (1.1.9)*to*-or bare infinitives and Gerund C) | | | |
| | | (1.1.10)Auxiliaries | (1.1.10.1)*do* support(C) | | |
| | | | (1.1.10.2) Other | (1.1.10.2.1)Syntactic error (C) | |
| | | | | (1.1.10.2.2)Semantic error (C) | |
| | | (1.1.11)Spelling, Capitalisation and Punctuation | (1.1.11.1)Punctuation between main clause and subordinate clause (C) | | |
| | | | (1.1.11.2)Spelling (C) | | |
| | | | (1.1.11.3)Other (C) | | |
| | | (1.1.12)Vocabulary and Phrase | (1.1.12.1)Word coinage (C) | | |
| | | | (1.1.12.2)Inappropriate word and phrase (C) | | |
| | | | (1.1.12.3) Words that are literally translated from Korean, or phrases that are either ungrammatical or literally translated from Korean (C) | | |
| | | (1.1.13)Clauses that are either ungrammatical or literally translated from Korean (C) | | | |
| | | (1.1.14)Other grammatical errors | (1.1.14.1)PossessiveC) | | |
| | | | (1.1.14.2)Word order in a phrase (C) | | |
| | | | (1.1.14.3)Omission of subject in a finite clause (C) | | |
| | | | (1.1.14.4)Other(C) | | |
| | (1.2)Unintelligible (C) | (1.2.1)Errors in clause construction | (1.2.1.1)Due to serious syntactic error (C) | | |
| | | | (1.2.1.2)Due to errors in clause construction, resulting in ambiguous and unclear clause (C) | | |
| | | (1.2.2)Use of unintelligible vocabulary (C) | | | |
| | | (1.2.3)Other (C) | | | |
| (2) Fluency | (2.1)Quantity (T) | (2.1.1)Less than 33% of (implicit) minimum quantity of around 200 words (i.e., 66 words) | | | |
| | | (2.1.2)Between 33% and 75% (i.e., 66 to 150 words) | | | |
| | | (2.1.3)Around 100% (i.e., 200 words) | | | |
| | | (2.1.4)More than 150% (i.e., more than 300 words) | | | |
| | (2.2)Coherence | (2.2.1)Disconnected and incoherent sentences for more than 50% of the whole script (T) | | | |
| | | (2.2.2) Local lack of coherence | (2.2.2.1) For an individual sentence | (2.2.2.1.1)Due to insufficient language command (C) | |
| | | | | (2.2.2.1.2)Due to its irrelevance to the previous sentence (C) | |
| | | | | (2.2.2.1.3)Due to its unintelligibility (C) | |
| | | | (2.2.2.2) For more than two consecutive sentences | (2.2.2.2.1)Due to inappropriate alignment of the sentences C) | |
| | | | | (2.2.2.2.2)Due to their irrelevance to the previous sentence (C) | |
| | | | | (2.2.2.2.3)Due to their unintelligibility (C) | |
| | (2.3)Cohesive devices | (2.3.1)Little repetitions using the substitution words (T) | | | |
| | | (2.3.2)Smooth connection between sentences using cohesive devices well (T) | | | |
| | | (2.3.3)Use of advanced connectors (C) | | | |
| | | (2.3.4)Errors in the use of number and person of pronouns (C) | | | |
| | (2.4)Advanced language (language level) | (2.4.1)English-like vocabulary, phrase and lexical phrases (T) | | (2.4.1.1)Uses them once or twice | |
| | | | | (2.4.1.2)Uses them more than three times | |
| | | (2.4.2)Good clause construction and good expansion of clauses through fluent use of adjective/adverbial clauses (T) | | (2.4.2.1)Uses them once or twice | |
| | | | | (2.4.2.2)Uses them across the script | |
| | | (2.4.3)Advanced grammar (C) | | (2.4.3.1)Use of complex aspect (C) | |
| | | | | (2.4.3.2)Use of relative adverbs (C) | |
| | | | | (2.4.3.3)Use of *that* clause for complement and subject (C) | |
| | | | | (2.4.3.4)Other(C) | |
| | | (2.4.4)Multi-word verb phrases (C) | | | |

(continued)

| | | | | | |
|---|---|---|---|---|---|
| (3) Organisation | (3.1)Paragraphing (T) | (3.1.1)No paragraphing | | | |
| | | (3.1.2)Errors in paragraphing (T) &(C) | | | |
| | | (3.1.3)Exact paragraphing | | | |
| | (3.2)Genre format and development | (3.2.1) Opening | (3.2.1.1)Blurred distinction between opening and body or lack of opening (T) | | |
| | | | (3.2.1.2)Genre format (T) | (3.2.1.2.1)Not follows the genre format | |
| | | | | (3.2.1.2.2)Follows the genre format | |
| | | | (3.2.1.3)Development (quantity) (T) | (3.2.1.3.1)Less than two sentences | |
| | | | | (3.2.1.3.2)Between two and three sentences | |
| | | | | (3.2.1.3.3)More than three sentences | |
| | | (3.2.2) Body | (3.2.2.1)Number of points (C) | | |
| | | | (3.2.2.2)Number of insufficiently developed points (C) | | |
| | | (3.2.3) Closing | (3.2.3.1)No closing (T) | | |
| | | | (3.2.3.2)Genre format (T) | (3.2.3.2.1)Not follows the genre format | |
| | | | | (3.2.3.2.2)Follows the genre format | |
| | | | (3.2.3.3)Development (quantity) | (3.2.3.3.1)Less than two sentences | |
| | | | | (3.2.3.3.2)Between two and three sentences | |
| | | | | (3.2.3.3.3)2More than three sentences, but not rounded off | |
| | | | | (3.2.3.3.4)More than three sentences and reasonably rounded off | |
| | (3.3)Topic address (T) | (3.3.1)The purpose/topic of the discourse is not explicitly addressed in the opening stage of the discourse | | | |
| | | (3.3.2)The purpose/topic of the discourse is not signalled in the opening stage of the discourse | | | |
| | | (3.3.3)The purpose/topic of the discourse is wrongly addressed in the opening stage of the discourse | | | |
| | | (3.3.4)The purpose/topic of the discourse is partly addressed in the opening stage of the discourse and the required topic is fully dealt with in the discourse or vice versa | | | |
| | | (3.3.5)The purpose/topic of the discourse is appropriately addressed in the opening stage of the discourse | | | |
| | (3.4)Content (T) | (3.4.1)The discourse omits some of the required content from the prompt | | | |
| | | (3.4.2)The discourse includes irrelevant points to the topic (T) &(C) | | | |
| | | (3.4.3)The discourse includes required content, but extremely simply and insufficiently | | | |
| | | (3.4.4)The discourse sufficiently includes required content | | | |

## Appendix 9. An example of a script coded according to the coding scheme

ID: 3-A0117B

Hi. I heard you want to travel my country, Korea. So, I'll give you advice about good places to visit. First, why don't you go to Kyung-ju? There are many Buddhist temples. You'll experience Korean traditional cultural properties. The Sogguram cave and Bulguksa temple are the most famous and important places. Though those were made about 1500 years ago. Both are still strong. Maybe you think Korean traditional construction skills are very special.

Second, Our old places are good, too. Kyungbokgoong is the main place. It is very huge and beautiful. Not only Korea people but also foreign sightseers love this place. In addition, This palace was made 500 years ago. Unfortunately, It experienced lots of fire accidents. But our ancestors fixed again and again. So It keeps itself safely.

If you want to know Korean young people, How about go to Myung-dong or Dae-hak lo? You can see the young's culture at Game Station, theatre, fashion center and even restaurants in Myung-dong. And COEX, which has a lot of entertainment systems. It is known by Mega-box and Mega web station, which young guys usually go. You'll experience what is "Modern Korea".

Last, What do you think about to visit agricultural villages? Frankly speaking, Cities are dirty and crowded. If you want to get fresh air and peace of mind, country give you calmness. Also, country people are very kind. If you are hungry, they'll set the table with chicken and Makgulri (Korean traditional wine) for you. And you can feel human love from old people. I'm sure they never turn down foreign tourists.

These are all what I can tell you. Some people say Korea is unmannered and a backward nation. Those are all lies. Don't worry you can't Korea! I'm certain that Almost all people help you kindly and minutely. If you don't trust my words, Please go to where I said! You'll never repent. So, Have a good time in Korea~!

**Appendix 10. Questionnaire III**

## Questionnaire on the first version of the rating scale

Dear _____,

Thank you very much for helping with the development of this rating scale.

Now that you have used the first version of the developmental rating scale, I would like to hear how you found the scale.

I would appreciate it if you would answer the questions below by either ticking where appropriate or writing your opinion.

Yours sincerely,

Jyi-yeon Yi
PhD candidate
University of Edinburgh

1.What was your initial impression when you first saw the rating scale? Please describe it.

2.About how long did it take you to fully understand the scale?

( ) Hour(s) ( ) minutes

3. About how long did it take you to rate a piece of writing using the rating scale?

( ) minutes

4.Was there any part which you had difficulty understanding?

(1) Yes

(2) No

5. If yes, please describe why you found this difficult.

6. If yes, please describe what it was.

7. Did you feel that any aspects of the rating scale were inconvenient to use?

(1)Yes

(2)No

8. If yes, please say why they were inconvenient.

9.Do you think that any part of the rating scale is unnecessary? If so, please write which one(s) is/are.

10.What do you think are the good points about the rating scale? Please describe them.

11.What do you think was different about using this rating scale and using the FCE rating scale?

Thank you very much for your co-operation.

**Appendix 11. Questionnaire IV**

## Questionnaire on the revised rating scale

Dear _____,

Thank you very much for helping with the development of this rating scale.

Now that you have used the revised rating scale, I would like to hear how you found it.

I would appreciate it if you would answer the questions below by either ticking where appropriate or writing your opinion.

Yours sincerely,

Jyi-yeon Yi
PhD candidate
University of Edinburgh

1.What was your initial impression when you first saw the rating scale? Please describe it.

2.About how long did it take you to fully understand the scale?

( ) Hour(s) ( ) minutes

3. About how long did it take you to rate a piece of writing using the rating scale?

( ) minutes

4.Was there any part which you had difficulty understanding?

(1)Yes

(2)No

5.If yes, please describe why you found this difficult.

6.If yes, please describe what it was.

7.Did you feel that any aspects of the rating scale were inconvenient to use?

(1)Yes

(2)No

8.If yes, please describe why they were inconvenient.

9. Do you think this rating scale is practical to use?

**Appendix 12. Questionnaire V**

## Questionnaire on the validity of the rating scale

Dear _____,

Now that you have used the newly developed rating scale for the English Writing course at foreign language high schools in Korea, I would like your opinion on its validity.
This questionnaire consists of nine open-ended questions. It would be greatly appreciated if you would answer each question in detail.

Jyi-yeon Yi
PhD candidate
The University of Edinburgh

1. Did you find that the manual on understanding and using the rating scale helped you to understand the rating scale? Was there any discrepancy between the description and categorisation in the manual and your knowledge of English that prevented you from understanding and using the RS2?

2. Do you think that you understood the descriptors in the RS2 as intended?

3. When did you consult the RS2 for your rating? For example, did you rarely use it during assessment? Did you only read it before you started rating? Having read it before starting the assessment, did you consult it whenever you needed to? Or did you consult it in order to verify your decision after you had chosen a band according to your own criteria?

4. Did you find any differences in applying the RS2 to assess scripts in the genre of a letter and scripts in the genre of a formal essay? For example, was it more applicable to the genre of letter than the genre of formal essay?

5. What are the advantages and disadvantages of using the RS2, compared with those of assessing according to your own subjective criteria?

6. Do you think that the using the RS2 would give you a valid picture of a student's writing ability?

7. This scale aims to assess writing ability in terms of accuracy, fluency and organisation. Do you think the scale actually does this?

8. This scale aims to assess writing ability in terms of accuracy, fluency and organisation. Did you consider any other aspects apart from these while you were assessing writing samples using the RS2? If so, what were they? Why do you think you considered these additional aspects? How did you deal with this situation?

9. When the RS2 is used to assess writing by Korean students, what impact /consequence do you think it will have on their writing ability and on the teaching and learning of writing? How do you think this compares with the impact / consequence of using the FCE rating scale?

Thank you very much for your co-operation

**Appendix 13. Questionnaire VI**

## Questionnaire on the Test-takers' Perception of the Rating Scale

**Tips for responding**

Please choose one which is the most appropriate among the given choices and circle it.

**I. Questions on the Respondent**

**Q 1**. Age

(1)14

(2)15

(3)16

(4)17

(5)over 18

**Q 2**. Gender

(1) male

(2) female

**Q 3**. Average English score in examinations at school

(1) below 50

(2) between 51 and 70

(3) between 71 and 80

(4) between 81 and 90

(5) between 91 and 100

**II. Questions on the Rating scale**

**Q 4**. In general, do you think the rating scale is appropriate as a rating scale for assessing Korean students' writing?

(1) Strongly disagree

(2) Slightly disagree

(3) Slightly agree

(4) Strongly agree

**Q 5**. Please put the reasons for your answer for Q 4.

**Q 6.** Does this rating scale appear to assess features which are appropriate for a rating scale for writing assessment?

(1) Strongly disagree

(2) Slightly disagree

(3) Slightly agree

(4) Strongly agree

**Q 7.** Please put the reasons for your answer for Q 6.

**Q 8.** Which feature(s) is/are included in the scale which you find unnecessary? Or which feature(s) is/are not included in the scale which you find necessary and should be included?

**Q 9.** Do you find the demarcation between bands appropriate?

(1) Strongly disagree

(2) Slightly disagree

(3) Slightly agree

(4) Strongly agree

**Q 10.** Please give the reasons for your answer to Q 9.

**Q 11.** Do you think this rating scale would provide a valid picture of your writing ability with your teacher?

(1) Strongly disagree

(2) Slightly disagree

(3) Slightly agree

(4) Strongly agree

**Q 12.** Please write the reasons for your answer to Q 11.

Thank you very much for your co-operation.